Khan and Martini *EURASIP Journal on Wireless Communications and Networking* (2016) 2016:93
DOI 10.1186/s13638-016-0584-6

EURASIP Journal on Wireless Communications and Networking
a SpringerOpen Journal

RESEARCH Open Access

QoE-driven multi-user scheduling and rate adaptation with reduced cross-layer signaling for scalable video streaming over LTE wireless systems

Nabeel Khan* and Maria G. Martini

Abstract

The scarcity of the available radio spectrum coupled with the growing popularity of bandwidth intensive mobile video applications poses a huge challenge to network operators. The solution of over-provisioning the network is not economical; hence, an appropriate strategy for scheduling and resource allocation among the users in the system is of crucial importance. This work focuses on scheduling multiple video flows on the downlink of a wireless system based on orthogonal frequency division multiple access (OFDMA), such as Long-Term Evolution (LTE) and LTE-A (LTE-Advanced) standards. We propose a joint multi-user scheduling and multi-user rate adaptation strategy providing an appropriate trade-off between efficiency and fairness, while ensuring high quality of experience (QoE) for the end users. We consider Scalable Video Coding (SVC) which facilitates the truncation of bit streams, thus allowing graceful degradation of video quality in the event of wireless channel variations or network congestion. The proposed scheduler utilizes QoE-aware priority marking, where video layers are mapped to priority classes and targets at minimizing delay bound violations for the most important priority classes under congestion. In order to reduce congestion, we propose multi-user rate adaptation at the MAC layer via a novel dynamic filtering policy for QoE-based priority classes.

Simulation results show that the proposed approach delivers to the end users a similar QoE as delivered by the state-of-the-art cross-layer approaches, where extensive cross-layer signaling, additional video rate adaptation modules at the core network, and frequent link probing from the wireless access network to the rate adaptation modules are required. The latter approaches are not implemented in real systems due to the aforementioned drawbacks, while our approach can be implemented without major modifications in the standard behavior of existing networks and equipment. The proposed framework can deliver delay-sensitive traffic as well as delay-tolerant best-effort traffic.

1 Introduction

The 4th-generation wireless technologies such as the 3rd Generation Partnership Project (3GPP) LTE/LTE-A and the enhanced capabilities of the recent smartphones and tablets have fostered the growth of multimedia and interactive bandwidth demanding services, such as live video streaming, video on demand, interactive gaming, and 2D and 3D video streaming over wireless networks. The Cisco Visual Networking Index (VNI) projects that video consumption will amount to 90 % of the global consumer traffic by 2019 [1]. However, supporting multimedia applications and services over wireless access, for instance, LTE base station (eNodeB) for LTE networks, is challenging due to constraints such as limited bandwidth and random time-varying channel conditions. The eNodeB is congested when the video traffic rate is higher than the wireless channel capacity. Scalable video, H.264/SVC [2] or Scalable High Efficiency Video Coding (SHVC) [3], is an attractive solution for real-time rate adaptation at the wireless access network. A scalable video stream has a base layer and several enhancement layers. As long as the base layer is received, the receiver can decode the video stream. As more enhancement layers are received,

*Correspondence: nabeelkhan48@yahoo.com
School of Engineering and Computing, Kingston University London, Penrhyn Road, KT1 2EE London, UK

the decoded video quality is improved. The enhancement layers can be dropped dynamically to match the wireless channel capacity.

While layer dropping provides a flexible way to perform rate adaptation, it also influences the user's perceived video quality. Customer satisfaction is the main objective for mobile network operators. Quality of experience (QoE) [4] reflects the user's experience and satisfaction for the service used. QoE evaluation can be performed via subjective tests with the help of a panel of users, in order to obtain a mean opinion score (MOS) [5] which reflect the quality perceived by the observers. This reflects the features of the human perceptual system, as dependent on human observation. Since subjective tests are time demanding and costly, objective video metrics have been developed to estimate the user's perceived quality. These are mathematical-based metrics, ranging, for video quality assessment, from mean square error (MSE) [6] and peak signal-to-noise ratio (PSNR) [6] to structural similarity (SSIM) [7] and more complex metrics possibly better estimating the perceived quality [8].

Content awareness at the eNodeB requires the contribution of video packets to the objective video quality. Content-aware multi-user packet scheduling and rate adaptation at the eNodeB play a crucial role in determining the overall customer experience. In order to provide QoE-based video delivery, two classes of strategies exist in the literature. One of the solutions comprises a video quality-aware packet scheduling and radio resource allocation strategy as proposed in [9–13]. The information on the content of different video traffic flows is provided through cross-layer signaling to the eNodeB. The main goal of the strategy is to maximize the video quality of the streaming users under wireless channel and bandwidth constraints. This strategy requires video processing to extract content information and/or signaling to deliver video content information at the eNodeB. The other approach, proposed in [14–17], is the utilization of a content-blind packet scheduler, such as proportional fair (PF), modified largest weighted delay first (M-LWDF), or exponential proportional fair (EXP-PF), at the eNodeB where content-aware radio resource allocation is performed at a proxy located close to the eNodeB. In this approach, a rate adaptation module avoids congestion at the eNodeB. This approach requires additional modules, proxy and rate adaptation modules, and excessive cross-layer signaling.

Current LTE/LTE-A networks are not built for QoE-aware video delivery. The LTE/LTE-A has a hierarchical architecture, where video packets are first passed through Packet Data Network Gateway (P-GW), located in the core network. The P-GW transfers video packets to the respective target eNodeB, where radio resources are assigned to the video packets. The P-GW has application-specific information but it is unaware of the congestion status at each eNodeB. On the other hand, the eNodeB is aware of the channel capacity and congestion status of the radio cell but it has no application-specific information. The authors in [18] proposed a QoE-aware video delivery by considering the hierarchical architecture of LTE/LTE-A network. The authors proposed a QoE-aware video packet marking at the core network and QoE-aware packet dropping at the eNodeB. The marking scheme at the core network transforms the video content information into QoE-aware priority classes. Therefore, the strategy avoids complex video content processing information at the eNodeB. Furthermore, no rate shaping modules are required at the core or radio access network. However, the authors utilized a content-blind packet scheduler. In our preliminary work [19], we studied that content-aware packet scheduling plays a key role in improving the overall QoE of the mobile users. In this work, we propose a content-aware scheduling and congestion avoidance strategy by utilizing the QoE-aware packet marking. The main contribution of this article can be summarized as follows:

- We propose a novel QoE-based priority-aware scheduling rule. As a key difference from existing content-aware scheduling rules, such as in [9–13], the scheduler avoids video content processing capabilities at the eNodeB by utilizing the QoE-aware priority marking.
- We propose a novel QoE-based rate adaptation strategy. The proposed strategy quantifies congestion by considering video packets sojourn time in the queue at the eNodeB. Packet delay plays an important role in QoE-based video delivery. The rate adaptation policy proposed in [18] is delay-blind, which can have a significant impact on the QoE-based video delivery.
- We propose a joint operation of QoE-aware scheduler and rate adaptation strategy. The joint operation is in contrast to the rule in [18], where packet dropping is content-aware but packet scheduling is content-blind.
- The proposed scheduling strategy considers the service needs of other traffic classes such as delay-constrained video conferencing and delay-tolerant web browsing traffic, also referred as the best-effort traffic. Therefore, the proposed rule is not only video quality-aware but also traffic type-aware.

The remainder of the paper is organized as follows. Related work is discussed in Section 2. Section 3 presents the considered system model and problem statement. Section 4 presents the proposed packet priority scheduler (PPS). After a description of the scheduling metric and an analysis of the factors composing it, a congestion avoidance methodology (priority class filtering policy) is

presented in Section 5. Section 6 presents the proposed scheduling rule for the best-effort traffic. The section also discusses how a mixture of delay sensitive and best-effort traffic users can be supported simultaneously. Simulation scenario and the considered benchmark strategies are presented in Section 7. Finally, the proposed joint scheduling function and priority class filtering policy are evaluated in Section 8 along with the state-of-the-art benchmark rules. Concluding remarks appear in Section 9.

2 Related work

In the literature, several QoE-based video delivery approaches have been proposed. In our previous work [20], we divide these strategies into two classes, based on the required video processing capabilities at the eNodeB, and/or additional modules requirements at the radio access network (RAN). These classes are briefly discussed in the following sections.

2.1 Content-aware scheduling and resource allocation at the eNodeB

With content-aware scheduling approaches, the optimization goal is the maximization of the video quality subject to time-varying wireless capacity. For instance in [21, 22], the concept of incrementally additive distortion is used to determine the importance of video packets for each user's precoded non-scalable video stream. Essentially, the increase in distortion due to the loss of a video packet is a function of all the other video packets that are dependent on it and cannot be decoded if it is not sent. This information is used to drop video packets in the event of congestion over the wireless interface, beginning with the least important video packet. The authors in [13] further extended the concept of video packet importance and proposed a joint packet scheduling and subcarrier assignment for OFDMA systems. According to [13], subcarrier is allocated through a two-level search path. In the first step for each flow, the video packet contributing largest to the video quality is selected. In the second step, the priority of each flow on a subcarrier is computed by considering the channel gain of the subcarrier and the quality contribution of the packet selected in the first step. The subcarrier is assigned to the flow having the highest channel gain and the video packet contributing largest to the video quality. A similar approach is presented in [12] where a flow is prioritized based on the ratio of the packet's video quality contribution and the number of subcarriers required to schedule the packet. The packet of a flow contributing largest to the video quality and in possession of the least number of subcarriers is scheduled in priority. Similar distortion-based joint packet scheduling and subcarrier assignment policies are presented in [9–11]. The strategies in [9–13] drop the packets violating the preassigned delay bound. However, none of the strategies consider packet deadline in the scheduling decisions.

The aforementioned content-aware scheduling strategies as well as the strategies proposed in [23–28] require video content information, such as distortion (if the video packet is dropped or scheduled successfully), decoding deadline associated with each of the video packets, decoding dependence of a video packet, error concealment strategy at the receiver, and several other parameters. These algorithms require extensive cross-layer signaling as well as video processing information at the eNodeB. However, the eNodeB manages wireless resources but it does not have access to complex video content information. Therefore, such scheduling algorithms pose problems from an implementation point of view.

2.2 Proxy-based content-aware resource allocation

The second approach employs a QoE optimizer which performs radio resource allocation decision for the eNodeB. The optimizer estimates the available radio resources for video transmission by retrieving channel quality information from the eNodeB. The optimizer takes into account the content characteristics of the video streams and performs video rate adaptation according to the available radio resources at the eNodeB. For instance, the authors in [14, 15] proposed a QoE-based video delivery by introducing two modules located inside the RAN. The two modules are the traffic engineering and traffic management module. The main task of the traffic management module is to act as the downlink optimizer for resource allocation, whereas the main task of the traffic engineering module is to act as a controller for performing rate adaptation in the RAN. The traffic engineering module performs rate adaptation either based on packet dropping or transcoding. The authors in [14] proposed three objective functions at the optimizer. One of the objective functions is the maximization of the MOS-based utility, according to which the rate adaptation is done to maximize the mean MOS (mean user-perceived quality). According to the objective function, resources are first reserved to the users with good channel quality and low-rate demanding applications. The authors also proposed a max-min fairness-based objective function, where the main goal of the objective function is to allocate resources such that all the users get the same perceived quality. However, the authors do not propose any scheduling algorithm to be used in conjunction with the proposed cross-layer resource allocation frame work. The scheduling algorithm is important in determining the overall performance of an LTE system. The work done in [16, 17] jointly addresses resource allocation and rate adaptation for Scalable Video Coding (SVC) traffic. The authors proposed a proxy-based solution with limited information exchange between the application and the MAC layer. The

main goal of the proposed framework is to maximize the sum of the achievable rates subject to the minimization of the distortion difference among multiple video flows.

The proxy-based approach requires additional modules at RAN and regular link probing from eNodeB to these modules. This approach can significantly increase the complexity of the network and raise the capital expenditure (CAPEX) for network operators.

3 System model and problem statement

The considered system model is shown in Fig. 1. We consider QoE-based packet marking at the P-GW, whereas packet scheduling and rate adaptation is performed at the eNodeB as shown in Fig. 1. The subsequent sections discuss the considered QoE evaluation approach, QoE-based packing marking strategy and the considered system model at the eNodeB.

3.1 QoE evaluation

QoE is a subjective measure, and performing subjective tests for real-time network management is not feasible. Therefore, objective QoE models have been built according to the ITU recommendations (e.g., [29, 30]). These models estimate perceptual video quality measured better than the traditional video quality metrics, such as MSE or PSNR. In this paper, QoE estimation is performed objectively by employing Video Quality Metric (VQM) [31]. VQM is a full reference metric which estimates video quality in terms of DMOS, perceptual quality difference between the original and the degraded video. DMOS can be mapped to MOS which ranges from 1 (worst quality) to 5 (best quality). MOS values around 3 represent an acceptable video quality with slightly annoying artifacts. The computed MOS constitutes a utility function which is used in a marking algorithm discussed in the following section.

3.2 QoE-based packet marking at P-GW

We consider a video server generating a pre-encoded video traffic workload. Video is assumed to be encoded in different layers according to the SVC standard [2] and temporally organized in units which can be decoded independently from each other, each referred as group of pictures (GOP). Packet marking strategies for SVC, such as in [32, 33], do not allow to compare and prioritize layers of multiple videos having different quality and rate characteristics. Therefore, we employ the content-aware packet marking algorithm described in [34]. The packet marking strategy in [34] allows network operators to adapt multiple video streams having different video layers and diverse quality and rate characteristics. For instance, Fig. 2 illustrates the basic idea of the QoE-based packet marking strategy. According to the figure, the marking algorithm maps scalable layers of two video streams to priority classes. The priority classes are identified by the priority class index j. Assuming 8 priority classes in the system, then packets marked with index 1, i.e., $j = 1$ belong to the least important priority class: as the index increases, the priority class importance increases. The algorithm exploits the utility functions (MOS vs. bitrate) of the video streams and marks layers according to their bitrates and contribution towards the overall perceived video quality. The main goal of the marking is to achieve the maximum overall QoE, maximization of video quality, under the constraint of the available network resources.

Apart from the SVC standard, the algorithm can mark video packets of H.264/AVC as well as newly developed High Efficiency Video Coding (HEVC) standard [35] and its scalable extension. Similar to the H.264/AVC standard, the temporal scalability feature is enabled in HEVC with a hierarchical temporal prediction structure. In H.264/AVC, temporal frame rates are enhanced by adding temporal layers. However, in HEVC, temporal sub-layers

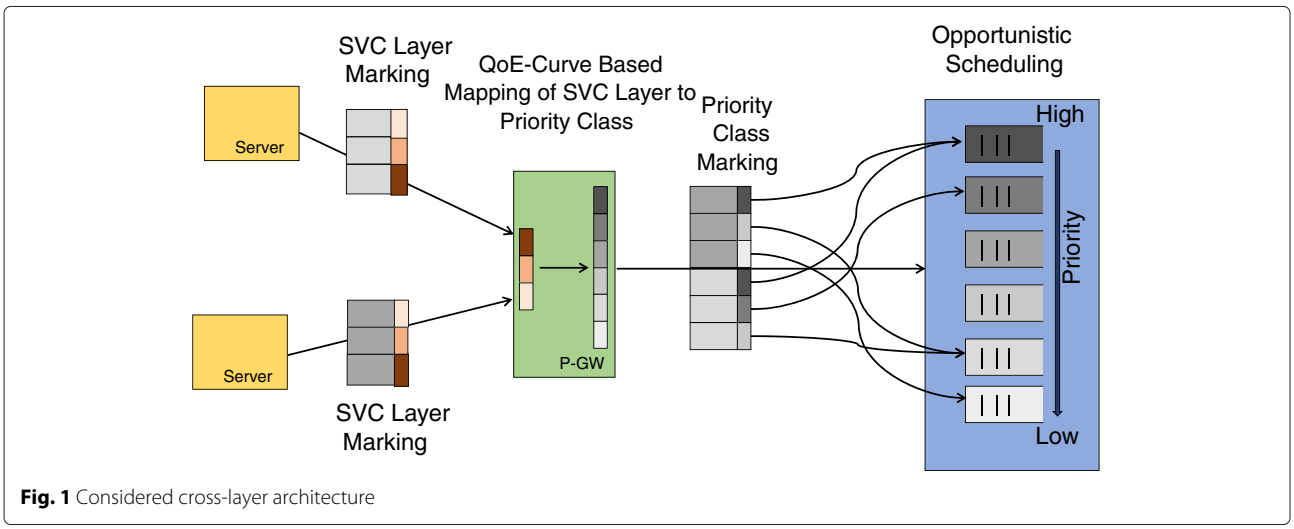

Fig. 1 Considered cross-layer architecture

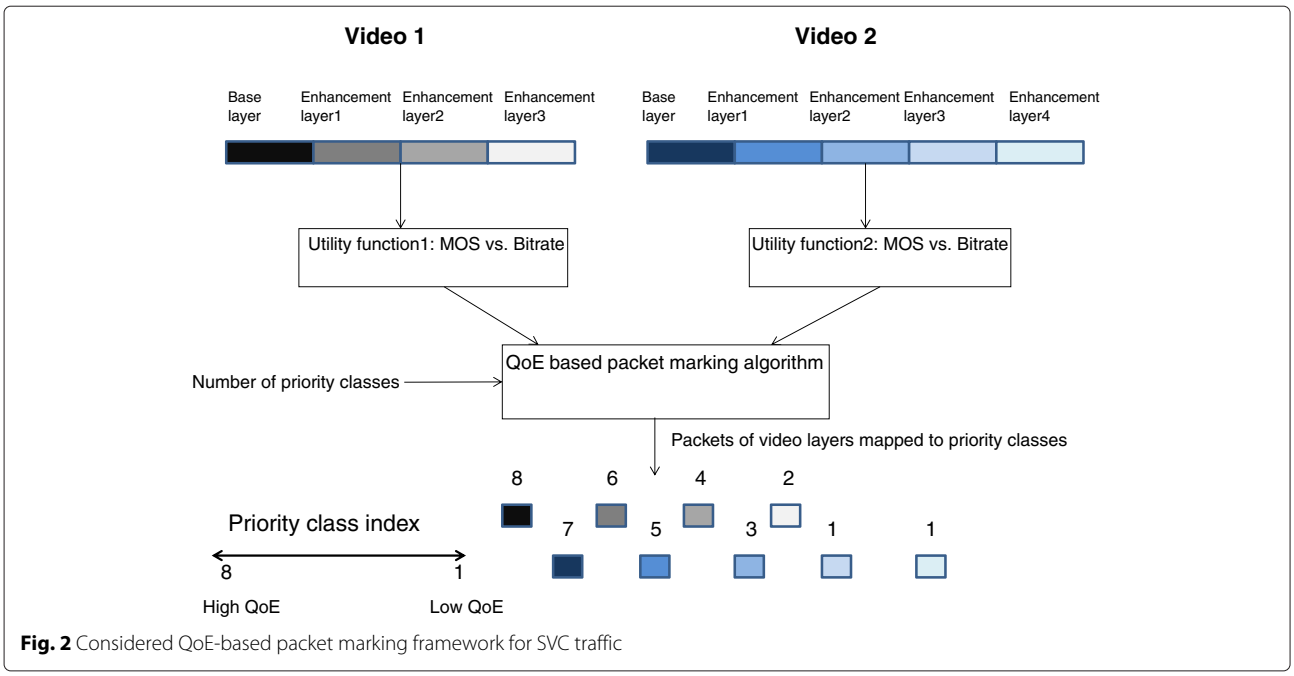

Fig. 2 Considered QoE-based packet marking framework for SVC traffic

corresponding to different frame rates are defined within a layer. Therefore, the marking algorithm can mark each temporal sub-layer according to its bitrate cost and QoE contribution. By reducing the frame rate, the transmission bitrate is adapted. Similar to SVC, Scalable HEVC (SHVC) provides quality as well as spatial scalability. Therefore, the utilization of content-dependent utility function, MOS vs. bitrate, enables packet marking of diverse video coding standards ranging from temporal scalable H.264/AVC to the scalable extension of HEVC.

3.3 System model at eNodeB

We consider a single-cell scenario in which the serving eNodeB is at the center of the cell. The serving eNodeB's medium access control (MAC) scheduler controls all the available Physical Resource Blocks (PRBs, the basic resource units in LTE) by allocating them to the active flows competing for resources.

Each user is assigned a queue at the eNodeB. The packet stream for each user at the eNodeB is referred to as a flow. Packets of a flow entering the buffer at the eNodeB are stored in first in first out (FIFO) order. It is important to note that SVC video streaming flows have QoE-based marked packets, identified by different priority class indexes, in their respective buffers. Furthermore, the packets of delay-sensitive priority classes entering the buffer are time stamped by the scheduler. The scheduler should assign enough resources to schedule the packets of these classes before the delay budget. Packets violating the delay budget are dropped from the queue since delay-sensitive traffic has no advantage in receiving expired packets.

In the cell, users report their instantaneous channel quality by means of a quantized feedback called Channel Quality Indicator (CQI). If according to the scheduling decision more than one PRB, with different CQI values on each PRB, are assigned to a user, then it is necessary to calculate an average supported CQI value. The scheduler uses the Mutual Information Effective SNR Mapping [36] method to calculate a single average CQI value to be used for all the allocated PRBs of a particular user. The single average CQI, assigned to a user, is transformed to number of bits according to the mapping reported in [37, 38].

3.4 Problem statement

In order to mathematically formulate the problem according to the information available at the eNodeB (priority class index, j, of each video packet and packet delay bound $D_{\max}$), we define the following parameters:

$H_i^{(n)}$: Head-of-line packet delay (waiting time of the packet residing in the buffer of a flow). The consideration of packet delay at the eNodeB is important since base and enhancement layers need to be scheduled before their respective decoding deadline at the receiver.

$\mathbb{1}_{i,j}$: Packet dependency indicator function for priority class j of flow i. This assumes value 0 or 1. The indicator function accounts for decoding dependency of video layers within the GOP.

$\sigma_{i,\varphi}^{(n)}$: An indicator whether PRB φ is used by flow i or not. This assumes value 0 or 1.

M_{PRB}: Total number of PRBs available for allocation at scheduling epoch n.

The objective of the strategy is to maximize the scheduling of packets, within the delay budget $D_{\max}$, with priority

$j_i^{(n)}$ over a moving average window of size t_w scheduling epochs:

$$\max\left(\sum_{m=n-t_w+1}^{n}\sum_{i=1}^{I}\mathbb{1}_{i,j}\cdot j_i^{(m)}\right) \tag{1}$$

subject to the following constraints

$$\begin{aligned}&\sigma_{i,\varphi}^{(n)}\in\{0,1\}\\&\sum_{i=1}^{I}\sigma_{i,\varphi}^{(n)}=1\end{aligned} \tag{2}$$

$$H_i^{(n)}\leq D_{\max} \tag{3}$$

The first constraint shows that each PRB can only be assigned to one flow at scheduling epoch n. The second constraint implies that each video packet must be scheduled before the delay bound $D_{\max}$, i.e., the HoL delay of a packet must be below the prescribed delay budget. A packet violating the preset HoL delay threshold is dropped from the buffer. The indicator function is 1 for packets of priority class j of flow i if all higher priority packets than class index j, over t_w scheduling epoch window, are successfully scheduled. On the other hand, the indicator function is 0 for packets of priority class j if any of the higher priority packets (packets marked with class index less than j), over the moving average window t_w, is dropped at the eNodeB.

The problem statement implies that, for video adaptation, packets are dropped from the highest enhancement layer to the first enhancement layer due to the decoding dependency (within the GOP) among the layers. The packets of the base layer need to be scheduled with the highest priority. Therefore, the joint scheduling and rate adaptation policy must ensure that a non-base layer should only be dropped when all of its higher enhancement layers are dropped.

The optimal solution of the above problem has been investigated in [12, 13], where packet importance is determined through the achieved distortion when the packet is successfully scheduled. According to [12, 13], after relaxing the non-linear and integer constraints, the optimal solution at each scheduling time interval requires $(\zeta\cdot I)^{M_{\text{PRB}}}$ computations, where ζ is the number of packets residing in the buffer of all the flows. However, according to [39], scheduling metrics requiring $I\cdot M_{\text{PRB}}$ computations can be implemented within a scheduling epoch of 1 ms (LTE's Transmission Time Interval). Furthermore, the work in [12, 13] requires distortion-based scheduling metric which requires extensive amount of video processing at the eNodeB.

4 Packet priority scheduler (PPS) for delay-sensitive priority classes

Packet scheduling is one of the most important functions of radio resource management (RRM) and plays a key role in distributing radio resources among different users with different service needs. It determines the overall performance of an LTE system. LTE is a multicarrier system where radio resources are spread in time and frequency domains. The basic time-frequency resource unit, called a PRB, is allocated to a user every 1-ms Transmission Time Interval (TTI). Defining a radio resource allocation on a per-PRB basis, as shown in Fig. 3, is simpler to implement and reduces complexity compared to complex strategies as proposed in [40, 41]. According to the figure, the user with the highest scheduling metric is allocated a PRB. With this approach, the scheduler computes $I\cdot M_{\text{PRB}}$ metrics per scheduling epoch. Therefore, the per-PRB scheduling rule has a linear dependence on the number of PRBs and flows.

Algorithm 1 shows the pseudo-code of the proposed PPS rule. PRBs are assigned in the increasing order of frequency. According to the pseudo-code, per-PRB scheduling metric $\Psi_{i,\varphi}^{(n)}$ is computed for each flow. PRB φ is assigned to flow i^* if it maximizes the scheduling metric $\Psi_{i,\varphi}^{(n)}$. The set of PRBs assigned to flow i^*, $\Phi_{\text{PRB},i^*}^{(n)}$, is updated on each PRB allocation.

Algorithm 1 Packet priority scheduler

```
repeat
  for φ = 1 to M_PRB do
    for i = 1 to I do
      Calculate Ψ_{i,φ}^{(n)}
    end for
    i* = argmax(Ψ_{i,φ}^{(n)})
    Assign PRB φ to flow i*
    Update Φ_{PRB,i*}^{(n)}
  end for
  n = n + 1
until END OF SIMULATION
```

The selection of the scheduling metric depends upon the desired performance requirements of the network operators which can be spectral efficiency (bit/s/Hz), fairness (guaranteeing a minimum performance threshold), or QoS provisions (packet delivery delay bound). It is important to note that the main business objective of mobile operators is the provision of multiplay applications of streaming and live video, VoIP and data on a single IP-based infrastructure. Therefore, the design of a QoS-aware scheduling strategy becomes mandatory. According to [39], QoS-aware scheduling strategies such as M-LWDF [42], EXP-PF [43], Log-rule [44], Exp-rules

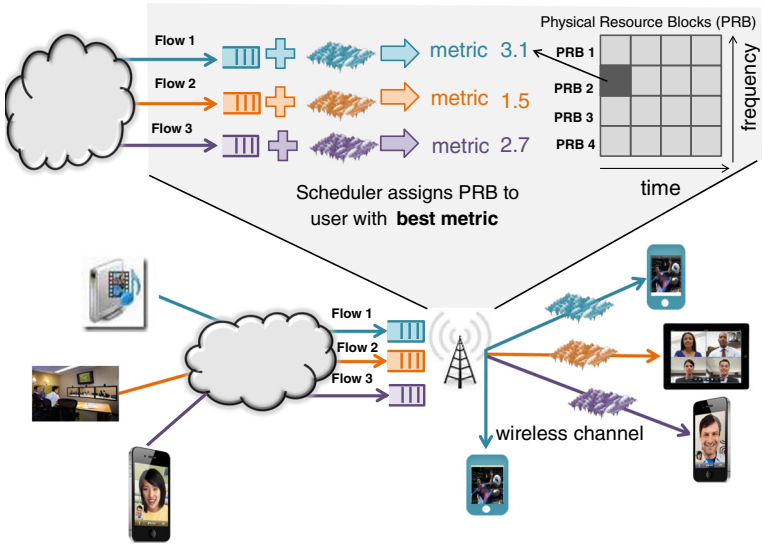

Fig. 3 Scheduling interactive and streaming video users. The scheduling metric must consider the different service needs of both types of video applications

[44], and other delay-based rules proposed in [45–47] are capable of meeting end users' flow requirements in terms of packet delivery delay bounds. In our considered framework, there are QoE-based marked priority classes for video streaming traffic, apart from different traffic types (real-time video and best-effort traffic). Therefore, a scheduling metric is required with a property of performing rate adaptation according to the importance of priority classes under the event of congestion. QoS-aware strategies in [39, 42–44] lack this important property. We propose a novel scheduling metric which considers the priority class index in the scheduling decisions and schedules the most important priority classes under congestion.

4.1 Scheduling metric

The main objective of the proposed scheduling function is to minimize the probability of delay bound violation of the most important packets by guaranteeing packet delivery before the delay bound. The HoL delay is defined as

$$H_i^{(n)} = n - n_{\text{enter}_i} \tag{4}$$

where n is the current scheduling epoch and n_{enter_i} is the scheduling epoch when the packet of flow i enters the buffer at the eNodeB. It is important to note that one of the most important QoS requirements is that packets have to be delivered within a delay bound. This constraint applies for QoE-based video streaming priority classes as well as for real-time applications such as video conferencing and VoIP. Each application has its own delay bound as reported in the LTE QoS Class Identifier QCI [48]. We propose to use HoL delay normalized by the target delay for packet delivery. The normalized HoL delay is given as

$$A_i^{(n)} = \frac{H_i^{(n)}}{D_{\max}} \tag{5}$$

where $D_{\max}$ is the delay bound of flow i's packets. In order to quantify the amount of congestion, we propose to use the HoL delay of all the flows in the system. The amount of congestion in the system is quantified by the normalized HoL delay averaged over the I delay-sensitive flows in the system and given as

$$\overline{A^{(n)}} = \frac{1}{I}\sum_{i=1}^{I} A_i^{(n)}. \tag{6}$$

The parameter $\overline{A^{(n)}}$ quantifies the system load: when $\overline{A^{(n)}}$ is equal to 0.5, it means that on average every flow's packet is experiencing an HoL delay of 50 % the delay bound. Due to diversity in the channel quality and data rate of the applications, some flows' packets may be experiencing a higher HoL and some less.

By considering the QoS constraints of different traffic types and QoE-based priority classes, we propose to use the following per-PRB scheduling metric:

$$\Psi_{i,\varphi}^{(n)} = W_{\varphi_i}^{(n)} \left[\frac{\chi_{i,\varphi}^{(n)}}{R_{i,\text{ave}}^{(n)}} \right] W_{q_i}^{(n)} \left[N_{q_i}^{(n)} \right] \tag{7}$$

The scheduling metric depends upon the following parameters:

- $N_{q_i}^{(n)}$ is the number of packets currently residing in the queue of flow i at scheduling epoch n. Considering the queue status in the scheduling decision avoids buffer overflow and keeps the queues stable.
- $W_{q_i}^{(n)}$ is the weight of the HoL packet. Mathematically, it is defined as:

$$W_{q_i}^{(n)} = \exp\left\{ \left[j_i^{(n)} \right]^{\overline{A^{(n)}}} A_i^{(n)} \right\} \tag{8}$$

where $j_i^{(n)}$ is the priority class index of the HoL packet of flow i. It is important to note that our proposed weight design depends on the system load. The higher the system load, the higher will be the normalized system delay $\overline{A^{(n)}}$, which results in a higher weight for the packets of the most important priority classes. If the system delay is low, packets from different priority classes have approximately the same weights. Therefore, the delay-based priority weight makes the scheduling rule dynamic. The impact of the exponential weight at different system loads (normalized system delay) and HoL delays is shown in Fig. 4. According to the figure, packets of priority classes 13, 7, and 1 have approximately the same weights when the normalized system delay is 0.25. Under congestion (normalized system delay of 0.75), the weight of the most important priority class increases exponentially, *w.r.t.* the lower priority classes, with the increase in packet's waiting time in the queue.

- $\chi_{i,\varphi}^{(n)}$ is the channel quality, in terms of bit/s/Hz, of PRB φ. It is important to note that effective utilization of the radio resources is extremely important. We utilize the CQI feedback from user i in determining the channel quality of PRB φ. This factor makes the scheduling rule channel aware and increases the system efficiency in terms of bit/s/Hz.
- $R_{i,\text{ave}}^{(n)}$ is the time-averaged throughput. Mathematically, it is defined as

$$R_{i,\text{ave}}^{(n)} = R_{i,\text{ave}}^{(n-1)} \left(1 - \frac{1}{n_w} \right) + \frac{1}{n_w} R_i^{(n-1)} \tag{9}$$

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

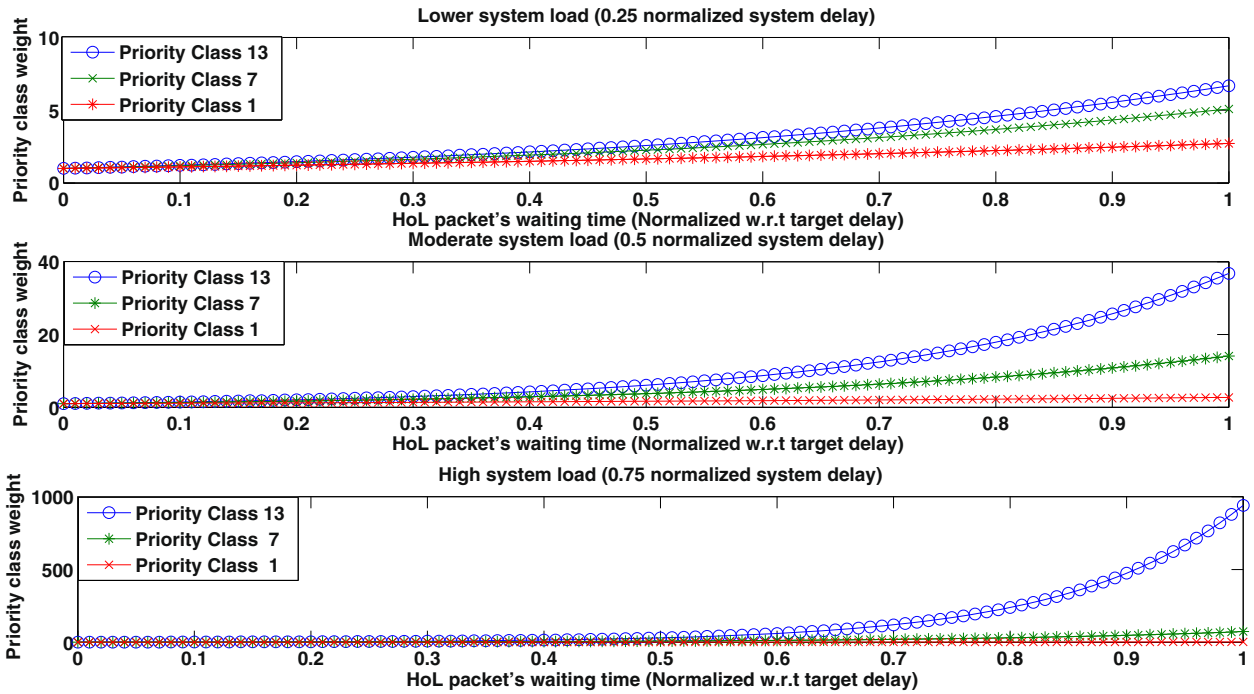

Fig. 4 Priority weight of the HoL packets (priority classes 13, 7, and 1) under normalized system delay of 0.25, 0.5, and 0.75

at scheduling instant $n-1$. n_w is the size of the time-average window. This term represents the achieved past average throughput by user i at scheduling instant n and is updated at every TTI. This provides proportional fairness in the scheduling decisions. The user experiencing the lower time-average throughput will be prioritized based on its channel conditions.

- $W_{\varphi_i}^{(n)}$ is the weight of the PRB.

$$W_{\varphi_i}^{(n)} = \frac{\chi_{i,\varphi}^{(n)}}{\chi_i^{(n)}} \tag{10}$$

where

$$\chi_i^{(n)} = \frac{1}{M_{\mathrm{PRB}}} \sum_{\varphi=1}^{M_{\mathrm{PRB}}} \chi_{i,\varphi}^{(n)}. \tag{11}$$

$\chi_i^{(n)}$ is the average PRB spectral efficiency of user i at scheduling instant n. $W_{\varphi_i}^{(n)}$ gives information on the variable amount of fading on the PRBs of each user. For instance, the reader can refer to Fig. 3. When the scheduler calculates the scheduling metric for PRB 2, the channel quality of the other 3 PRBs (PRB 1, PRB 3, and PRB 4) of the user is not considered. However, with $W_{\varphi_i}^{(n)}$, the channel quality of all the PRBs is considered in the scheduling metric. If a user is experiencing a high interference on some of the PRBs and other PRBs have better channel quality, then this factor assigns a lower weight to the PRBs with poor channel quality. On the other hand, the PRBs with the best channel quality for a user will be assigned a higher weight, thus utilizing the independent multi-user frequency selective fading.

At moderate normalized system delay, the scheduler acts like a queue-aware proportional fair scheduler, as the priority class weights are approximately the same. Therefore, the channel quality of a flow has a higher impact in the scheduling decisions. At higher normalized system delay, the weight function increases exponentially for the most important priority classes. Therefore, the scheduler minimizes the delay bound violations of packets from the most important priority classes. Under congestion, the scheduler acts like a strict priority scheduler with less importance of channel quality and more importance of higher priority class packets in the scheduling decisions.

The next section discusses a novel congestion avoidance methodology at the MAC layer which avoids the overloading of the scheduling function.

5 Congestion avoidance through dynamic priority class filtering

QoE-based SVC layer priority marking prioritizes multiple video streams based on their QoE contribution. Rate adaptation on QoE-based priority classes has the potential of achieving optimal radio resource utilization. Unlike state-of-the-art scheduling strategies, the scheduler proposed in Section 4 exploits packet marking and rate adaptation (under congestion) is performed by prioritizing the important priority classes and reducing the resource allocation probability of the less important priority classes. However, rate adaptation solely relying on the scheduling function presents the following issues:

- Under congestion, the less important packets residing in the buffer till the delay bound block packets of important priority classes. This phenomenon is also known as HoL blocking. When the HoL packets belong to the least important priority class then, under congestion, higher priority packets have to wait to be scheduled till the least important priority class packets are dropped from the buffer.
- At high system delay, packets belonging to low importance priority classes residing in the buffer at eNodeB are dropped when their delay bound is reached. Lower priority class packets, residing in the buffer till the delay bound, increase the average system delay, which in turn increases the resource allocation probability of the most important priority classes. The system becomes strictly priority driven when the normalized average system delay is high, i.e., high system delay reduces the channel awareness of the scheduler: highest priority packets are assigned resources with reduced importance of channel quality in the scheduling decisions. This leads to a reduction in the system efficiency (bit/s/Hz).
- The resource allocation probability of less important priority classes depends upon the system load (normalized system delay). However, due to the probabilistic arrival of the incoming traffic and stochastic nature of the wireless channel, the resource allocation probability of less important priority classes changes randomly. This leads to fluctuation in the perceived video quality of the users due to variation in the system load and wireless channel capacity. According to [49], multimedia services users would usually prefer to keep a fairly constant quality level rather than being exposed to fluctuations in the video quality.

In light of the aforementioned issues, we propose a novel concept of multi-user rate adaptation by applying a filtering policy on QoE-based priority classes of SVC flows. The joint operation of the scheduler and multi-user rate adaptation is shown in Fig. 5. The priority class filter utilizes the metrics calculated by the scheduling function and blocks the least important priority classes from entering the buffers at the eNodeB. Under congestion, the PPS scheduling function reduces

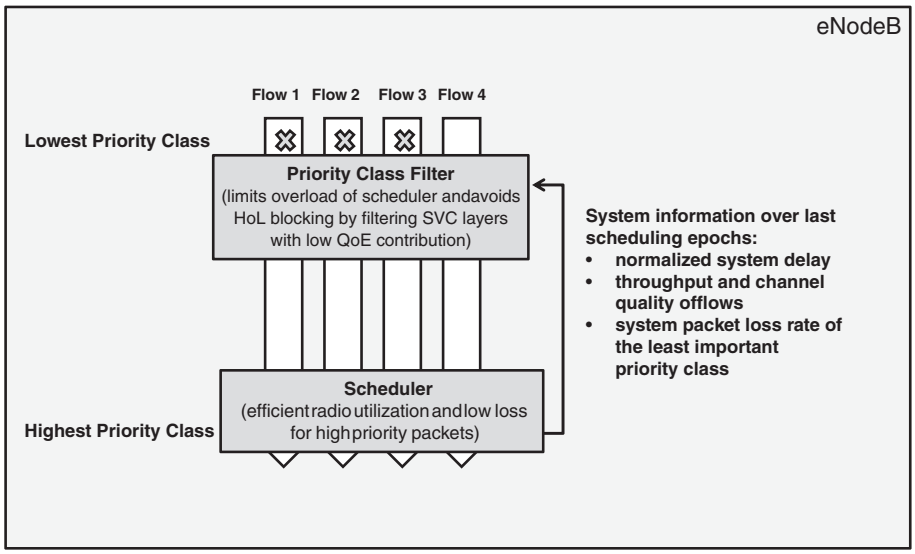

Fig. 5 Interaction between the scheduler and priority class filter

the resource allocation probability of the least important priority classes. The result is a delay bound violation of lower priority packets. The priority class filtering algorithm exploits these characteristics of the scheduling function and reduces congestion and HoL blocking. Furthermore, it also reduces fluctuations and provides fairly constant perceived video quality. The details of the priority class filtering algorithm are provided in the following section.

5.1 Hysteresis-based policy for rate adaptation through dynamic priority class filtering

In order to quantify the congestion, the PPS scheduling function calculates the normalized average system delay $\overline{A^{(n)}}$ at each scheduling epoch. We propose to utilize this congestion information in the priority class filtering decisions. Figure 6 shows the basic concept of the proposed policy. According to the figure, there is an upper and lower limit of the normalized average system delay. The proposed window-based filter policy is based on dual threshold action. The decision whether a flow's priority class packet is allowed or blocked from entering the buffer is taken every t_w scheduling epochs. Restricting the admission of priority class packets under higher system delay (point P1 in Fig. 6) will decrease the average system delay and increase the scheduler's channel awareness, resulting in better exploitation of multi-user channel diversity as shown in point P2. In order to reduce the under utilization of radio resources, the re-admission of priority classes is very important as shown in point P3. Therefore, the main goal of the filtering policy is to maintain the average system delay between the two thresholds.

We propose to perform rate adaptation by dropping packets marked with the lowest priority class index (less important priority class in terms of user satisfaction). Let j^* be the lowest priority class index, $\theta(i,j^*)$ be the set of flows having packets of priority class j^*, and $\delta(i,j^*)$ be the admission control vector containing the *IDs* of the blocked flows for priority class j^*. Flows' priority class j^* packets are blocked from entering their respective buffers if their *IDs* are removed from $\theta(i,j^*)$ and added to $\delta(i,j^*)$. When flows' priority class j^* packets are re-admitted, then their *IDs* are transferred back from $\delta(i,j^*)$ to $\theta(i,j^*)$. The filtering process comprises a two-step policy. Step one computes the number of flows to block or re-admit for the least important priority class j^*, whereas step two identifies the flows whose priority class j^* is blocked from entering the buffer at eNodeB.

5.1.1 Step one

According to the PPS scheduling metric, the priority weight decreases the resource allocation probability of the lowest importance priority class when the normalized instantaneous system delay is high. In order to facilitate the priority class filter decisions, we calculate the number of transmitted packets of the current lowest priority class represented by index j^* and the number of packets dropped due to delay bound violation. Let $n^{'}$ be the scheduling epoch when a priority class filtering decision is taken. The system packet loss ratio, over t_w scheduling epochs, at $n^{'}$ is:

$$\mathrm{plr}_{j^*}^{(n^{'})} = \frac{\sum_{m=n^{'}-t_w+1}^{n^{'}} P_{\mathrm{drop}}^{(m)}}{\sum_{m=n^{'}-t_w+1}^{n^{'}} \left(P_{\mathrm{transmit}_{j^*}}^{(m)} + P_{\mathrm{drop}}^{(m)} \right)} \tag{12}$$

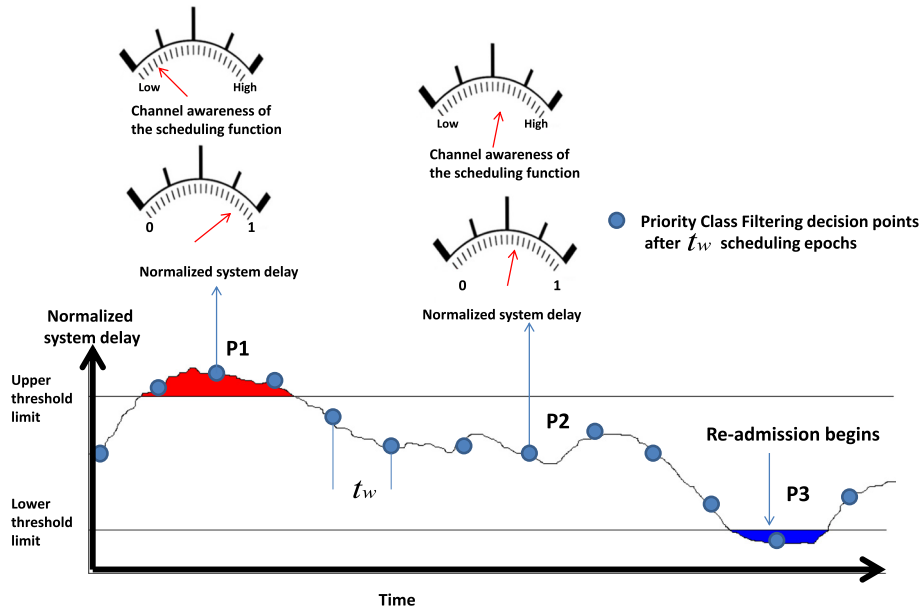

Fig. 6 Hysteresis principle for dynamic priority class filtering

where

$P^{(m)}_{\text{transmit}_{j^*}}$: Number of transmitted packets of class j^* over the moving average transmission window t_w.

$P^{(m)}_{\text{drop}}$: Number of dropped packets over the moving average transmission window t_w.

The scheduling metric computes the normalized system delay, $\overline{A^{(n)}}$, which quantifies the congestion status of the network. We propose to utilize the normalized system delay averaged over t_w scheduling epochs:

$$H^{(n')} = \frac{1}{t_w} \sum_{m=n'-t_w+1}^{n'} \left(\overline{A^{(m)}}\right) \tag{13}$$

where $H^{(n')}$ indicates congestion in the network by calculating the average of the normalized system delay over the moving average transmission window of size t_w epochs. Flows are blocked or re-admitted according to the following rules:

$$N_{\text{block}_{j^*}} = \left\lfloor I_{j^*} \cdot H^{(n')} \cdot S_{\text{hyst}} \cdot \text{plr}^{(n')}_{j^*} \right\rfloor \tag{14}$$

$$N_{\text{re-admit}_{j^*}} = \left\lfloor I_{j^*} \cdot \left(1 - H^{(n')}\right) \cdot \left(1 - S_{\text{hyst}}\right) \right\rfloor \tag{15}$$

where $N_{\text{block}_{j^*}}$ is the number of flows blocked for class j^* and $N_{\text{re-admit}_{j^*}}$ is the number of flows re-admitted. The number of flows to block or re-admit is taken once every t_w scheduling epochs. S_{hyst} is the output of the hysteresis-based window process. I_{j^*} denotes the total number of flows for priority class j^*. The higher the congestion, the higher the delay bound violations which in turn results in a higher number of flows blocked for priority class j^*. Similarly, the lower the normalized system delay, the higher the number of re-admitted flows. The hysteresis-based window output S_{hyst} adds stability in the blocking and re-admission decisions. It is based on the principle of dual threshold, which states that when the input is higher than a certain chosen threshold, the output is high. When the input is below a different (lower) chosen threshold, the output is low; and when the input is between the two levels, the output retains its last value. Mathematically, it is defined as:

$$S^{(n')}_{\text{hyst}} = \begin{cases} P\left(S_{\text{thr}_\text{h}}\right), & \text{if } H^{(n')} > S_{\text{thr}_\text{h}} \\ S^{(n'-t_w)}_{\text{hyst}}, & \text{if } S_{\text{thr}_\text{l}} \le H^{(n')} \le S_{\text{thr}_\text{h}} \\ \rho^{(n')}, & \text{if } H^{(n')} < S_{\text{thr}_\text{l}} \end{cases} \tag{16}$$

where

$$\rho^{(n')} = S^{(n'-t_w)}_{\text{hyst}} \cdot (1-\omega) + \omega \cdot P(S_{\text{thr}_\text{l}}) \tag{17}$$

S_{thr_l} and S_{thr_h} are the lower and higher threshold limits of the averaged normalized system delay, $H^{(n')}$, respectively. Similarly $P(S_{\text{thr}_\text{h}})$ and $P(S_{\text{thr}_\text{l}})$ are the probability of delay bound violation based on the congestion status of the network. The higher the congestion, the higher the probability of delay bound violation. Intuitively, the higher threshold limit S_{thr_h} of the normalized system delay is set to a point where the probability $P(S_{\text{thr}_\text{h}})$ of delay bound violation is 1. Similarly, the lower threshold limit is set to a point where the probability of delay bound violation is

very low. $\rho^{(n')}$ is a factor which determines urgency in the re-admission decisions, ω is a constant (less than 1) used in the exponential moving average weight, $\rho^{(n')}$. The speed at which the hysteresis output decreases, every t_w cycle, depends up on ω. The higher the ω (close to 1), the higher the re-admission speed (under low system delay) for priority class j^* flows.

The basic operation of the hysteresis based priority class filter is shown in Fig. 7. According to the figure, the output $S_{\text{hyst}}^{(n')}$ latches to $P(S_{\text{thr}_h})$ when the congestion parameter $H^{(n')}$ crosses the higher threshold limit. The hysteresis output is latched to $P(S_{\text{thr}_h})$ when $H^{(n')}$ is in between the two thresholds, i.e., the hysteresis output is retained to its last value $S_{\text{hyst}}^{(n'-t_w)}$. When the congestion parameter $H^{(n')}$ reaches the lower threshold limit as shown by the point P1, the output S_{hyst} changes to ρ which is the exponential moving average equation given in (17). The output decreases, after every t_w epochs, according to the moving average Eq. (17) shown by points P2 and P3. The hysteresis output is retained, $S_{\text{hyst}}^{(n'-t_w)}$, when the congestion parameter $H^{(n')}$ crossovers the lower threshold limit as shown by point P4. The hysteresis output remains latched until $H^{(n')}$ crosses either the lower or the higher threshold limit.

5.1.2 Step two

After computing the number of flows to block for priority class j^*, the next step is to determine the flows whose priority class j^* is blocked from entering the buffer at eNodeB. We propose to block the flows having the lowest ratio of channel quality to time-averaged throughput. In order to compute this, we utilize the average PRB spectral efficiency $\chi_i^{(n')}$ of user i computed by the scheduling function in (11). After t_w scheduling epochs, the time-averaged channel quality $\overline{\chi}_i^{(n')}$ is given as

$$\overline{\chi}_i^{(n')} = \frac{1}{\chi_{\max}} \left[\frac{1}{t_w} \sum_{k=n'-t_w+1}^{n} \chi_i^{(k)} \right]. \tag{18}$$

Furthermore, we also utilize the time-averaged throughput $R_{i,\text{ave}}^{(n)}$ computed by the scheduling function in (9). The metric $\alpha_i^{(n')}$ determines which of the flows have to be blocked or re-admitted and is given as

$$\alpha_i^{(n')} = \frac{\overline{\chi}_i^{(n')}}{R_{i,\text{ave}}^{(n)}}. \tag{19}$$

5.1.3 The pseudo-code of hysteresis based priority class filter policy

The pseudo-code of the priority class filtering process is shown in Algorithm 2. According to the algorithm, the

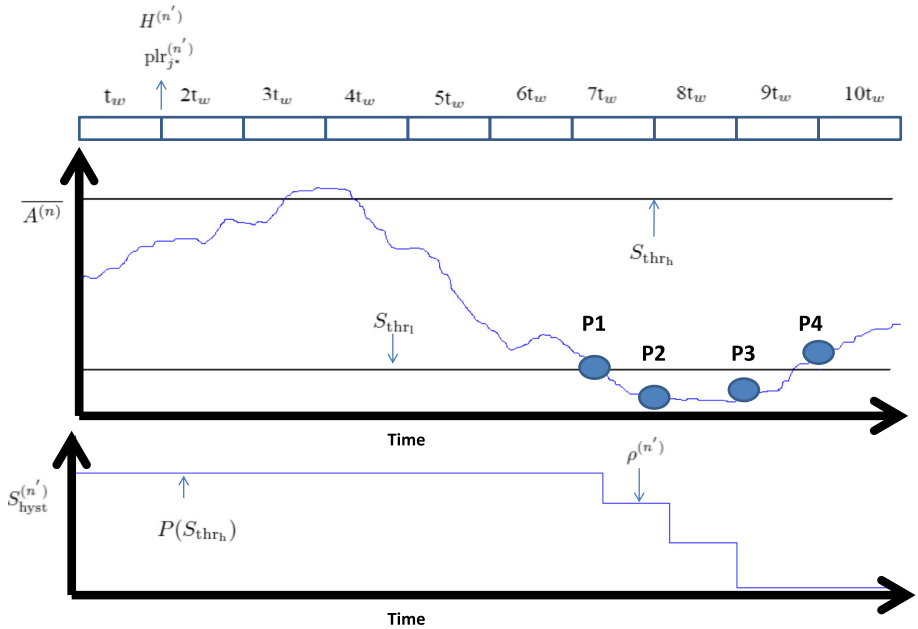

Fig. 7 Basic operation of hysteresis based priority class filter policy

Algorithm 2 Hysteresis based priority class filter policy

```
Set Simulation_time
repeat
  for m = n to m = n' do
    Calculate P^(m)_transmit_j* and P^(m)_drop at each scheduling epoch
  end for
  Calculate plr^(n')_j*, for class j*, according to (12)
  Calculate H^(n') according to (13)
  if plr^(n')_j* > 0 then
    For class j* calculate the number of flows to block, N_block_j*, according to (14)
    if N_block_j* > 0 then
      Update δ(i,j*) and θ(i,j*) by computing θ^asc(i,j*)
      if |θ(i,j*)| == 0 then
        j* = j* + 1
      end if
    end if
  else
    For class j* calculate the number of flows to re-admit, N_re−admit_j*, according to (15)
    if N_re−admit_j* > 0 and H^n' < S_thr_l then
      Update δ(i,j*) and θ(i,j*) by computing δ^desc(i,j*)
      if |δ(i,j*)| == 0 then
        j* = j* − 1
      end if
    end if
  end if
until END OF Simulation_time
```

number of transmitted packets of the current lowest priority class j^* and the number of packets dropped due to delay bound violations are calculated at each scheduling epoch. After t_w scheduling epochs, the packet loss ratio $\text{plr}_{j^*}^{\left(n'\right)}$ and congestion parameter $H^{\left(n'\right)}$ are computed according to (12) and (13), respectively. The filter blocks or re-admits flows based on the following conditions:

- If $N_{\text{block}_j^*} > 0$: The filter transforms the set of flows for priority class j^*, $\theta(i,j^*)$, to $\theta^{\text{asc}}(i,j^*)$. $\theta^{\text{asc}}(i,j^*)$ contains the set of flows sorted in an ascending order of $\alpha_i^{(n')}$. During the priority class filter decision epoch, the first $N_{\text{block}_j^*}$ *IDs* are removed from $\theta^{\text{asc}}(i,j^*)$ and added to $\delta(i,j^*)$. If priority class j^* for all the flows is blocked, then $|\theta(i,j^*)|$, cardinality of $\theta(i,j^*)$, is 0. If this condition is met, then, in the next filter cycle, packets of priority class j^*+1 are blocked based on the congestion and packet loss ratio. It is important to note that if the current lowest priority class is j^*, then all flow packets marked with index less then j^* are blocked from entering the buffer.
- If $N_{\text{re-admit}_j^*} > 0$ and $H^{(n')} < S_{\text{thr}_l}$: The admission control row vector $\delta(i,j^*)$ is transformed to $\delta^{\text{desc}}(i,j^*)$. $\delta^{\text{desc}}(i,j^*)$ contains the set of flows sorted in a descending order of $\alpha_i^{(n')}$. If $H^{(n')}$ is below S_{thr_l}, then S_{hyst} will decrease exponentially, according to (17), with each t_w cycle. S_{hyst} will continue to decrease until $H^{(n')} \geq S_{\text{thr}_l}$. After t_w scheduling epochs if $N_{\text{re-admit}_j^*} > 0$, then $\delta(i,j^*)$ is transformed to $\delta^{\text{desc}}(i,j^*)$. First, $N_{\text{re-admit}_j^*}$ number of flows' *IDs* are removed from $\delta^{\text{desc}}(i,j^*)$ and added to $\theta(i,j^*)$. If all the flows for priority class j^* are re-admitted then after t_w epochs, flows for priority class j^*-1 will be re-admitted based on $N_{\text{re-admit}_j^*}$.

5.2 Computational complexity

In this section, we analyze the computational complexity of the proposed strategy. Let I refer to the total number of flows in the system and M_{PRB} refer to the total number of PRBs in the system. The worst case computation complexity appears at the priority class filter decision epoch n' which consists of scheduling and priority class filtering policies. The scheduling rule

computes $I \cdot M_{\mathrm{PRB}}$ metrics and thus has a complexity of $O(I \cdot M_{\mathrm{PRB}})$.

The metrics such as normalized system delay $\overline{A^{(n)}}$, time-averaged throughput $R_{i,\mathrm{ave}}^{(n)}$, and channel quality $\overline{\chi}_i^{(n)}$ are computed by the scheduler at each scheduling epoch. Therefore, the processing burden introduced by the filter is minimal. The priority class filtering decision is performed once every t_w scheduling epochs. The first step computes the number of flows to block or re-admit. This step computes only one metric, either (14) or (15). The computation complexity for this step, $O(1)$, is independent of the number of flows and priority classes. The second step identifies the flows whose least important priority class is blocked or re-admitted. This step computes α for each of the flows. Furthermore, the admission control vector is sorted according to α. Therefore, the computation of α and sorting requires $O(I + I.\log(I))$ operations.

6 Scheduling rule for the best-effort traffic class

Delay-tolerant traffic (web browsing or data transfer applications such as FTP) is regarded as best-effort traffic class. QoS constraints, in terms of packet delay, of best-effort traffic class are not as stringent as that of video conferencing and video streaming applications. It is important to note that best-effort traffic corresponds to a significant proportion of mobile's traffic. It has been reported in [1] that by 2019, 72 % of total mobile's traffic will be video, followed by the Web/Data traffic contributing to 19 % of total mobile traffic. Therefore, delay-tolerant traffic needs to be prioritized carefully and must be part of an overall traffic shaping scheme. In order to consider delay-sensitive and delay-tolerant traffic classes simultaneously, the following criterion must be met:

- Delay-sensitive flows must meet their desired QoS, and delay-tolerant flows must receive the maximum possible throughput without compromising the QoS constraints of the delay-sensitive flows.

In the literature [39], composite scheduling rules serve best-effort traffic by using the classical proportional fair rule, i.e., ratio of instantaneous channel quality to the time-averaged throughput. They prioritize delay-sensitive traffic by considering either the logarithmic, exponential, or linear function of the HoL delay. In this work, we propose to dynamically prioritize best-effort traffic by utilizing the normalized system delay. We design a dynamic weight for the best-effort traffic which is $W_{\mathrm{best-effort}}^{(n)} = C^{1-\overline{A^{(n)}}}$ with $C > 1$, where $\overline{A^{(n)}}$ is the normalized system delay of the delay-sensitive flows and C is a prioritization factor for the best-effort traffic. The higher the prioritization factor, the higher the priority weight for the best-effort traffic under lower normalized system delay. The prioritization weight at different system delays and with different values of C is shown in Fig. 8. The composite scheduling rule for both the traffic types is

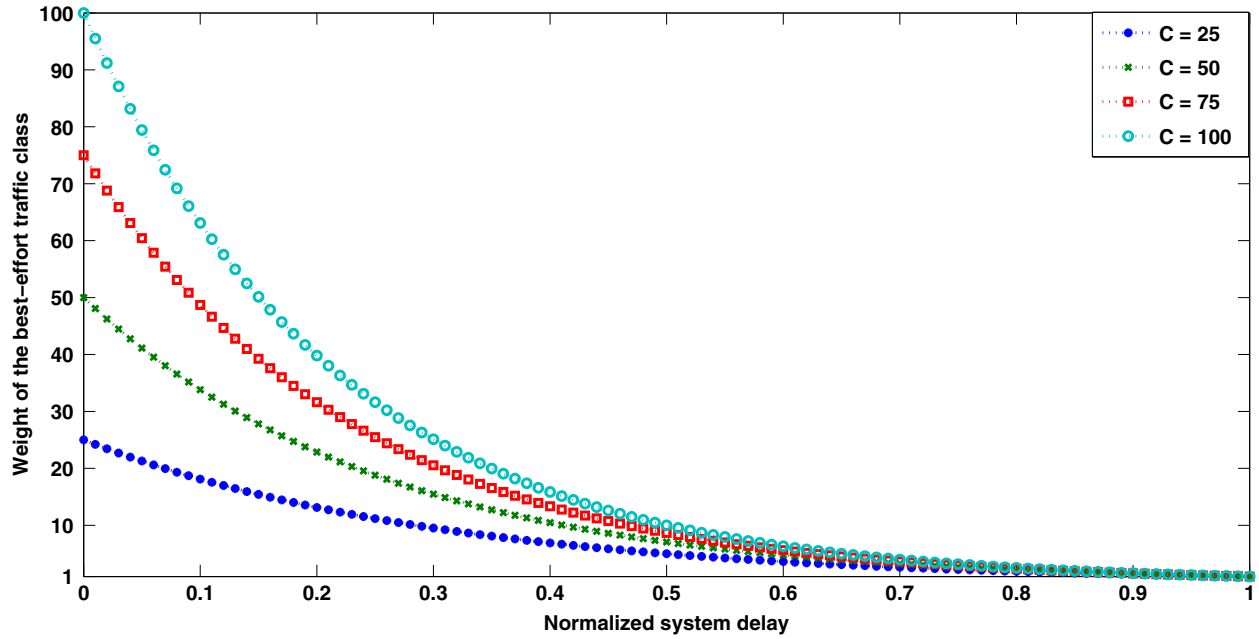

Fig. 8 Priority weight of the best-effort traffic class

$$\Psi_{i,\varphi}^{(n)} = \begin{cases} W_{\varphi_i}^{(n)} \left[\frac{\chi_{i,\varphi}^{(n)}}{R_{i,\text{ave}}^{(n)}} \right] W_{q_i}^{(n)} \left[N_{q_i}^{(n)} \right], & \text{if } j \in \text{delay-sensitive traffic classes} \\ W_{\text{best-effort}}^{(n)} \left[\frac{\chi_{i,\varphi}^{(n)}}{R_{i,\text{ave}}^{(n)}} \right] & \text{if } j \in \text{best-effort traffic class} \end{cases} \tag{20}$$

Delay-sensitive traffic classes (QoE-based classes for SVC traffic, video conferencing, and VoIP classes) are prioritized by the scheduling rule proposed in Section 4, whereas the priority of the best-effort traffic depends upon the normalized system delay of delay-sensitive flows. At higher normalized system delay, the scheduling rule for the best-effort traffic reduces to a simple proportional fair rule as the dynamic weight reduces to 1 as shown in Fig. 8. The aforementioned composite scheduling rules along with the hysteresis-based filter on QoE-based priority classes have three important properties:

- *Higher normalized system delay*: The delay-sensitive traffic classes are prioritized by the consideration of the exponential weight $W_{q_i}^{(n)}$ and the queue size $N_{q_i}^{(n)}$ in the scheduling decisions. Therefore, QoS constraints of the delay-sensitive traffic are always met under congestion. Under congestion, the backlog (packets waiting to get scheduled) of the best-effort traffic increases. When the system is heavily congested with delay-sensitive traffic, the filter blocks bandwidth demanding video traffic's lower priority classes which reduces the normalized system delay.
- *Moderate normalized system delay*: If the normalized system delay is between the two thresholds, the priority of the best-effort traffic class depends upon the number of packets residing in the buffers of delay-sensitive flows. The higher the number of packets residing in the buffer, the higher the probability of congestion which results in a higher resource allocation probability of the delay-sensitive flows. If the queue size of delay-sensitive flows is such that their input traffic rate is in the achievable rate region (the queue size remains stable), then the resource allocation probability of best-effort traffic increases. Under such condition, best-effort traffic flows get scheduled subject to the condition that the queue size of the delay-sensitive flows remains stable.
- *Lower normalized system delay*: Under such condition, the resource allocation probability of the best-effort traffic class is maximum as shown by the proposed weight design in Fig. 8. The result is a reduction in the backlog of the best-effort traffic, thus fully exploiting the variable bitrate characteristics of the video traffic as well as the probabilistic arrival of the incoming traffic.

In the subsequent sections, we analyze the performance of the QoE-based joint scheduling and priority class filtering strategy.

7 Simulation setup

7.1 Simulation scenario 1

In order to investigate the performance of the proposed joint scheduling and priority class filtering algorithm, an LTE link-level simulator [37, 38] built on MATLAB's object-oriented features is selected as the simulation platform. The video sequences are encoded with the SVC codec (medium grain scalability (MGS) [2]) and comprise a base layer and 12 quality layers. MGS scalability provides sufficient bitrate granularity for rate adaptation. The increase in MOS score, computed by VQM to MOS mapping as reported in Section 3.1, along with the addition of each quality layer is shown in Fig. 9. The wireless simulation parameters are reported in Table 1. According to the simulation parameters, the average system capacity is approximately 6.8 Mbps (2.266 bit/s/Hz

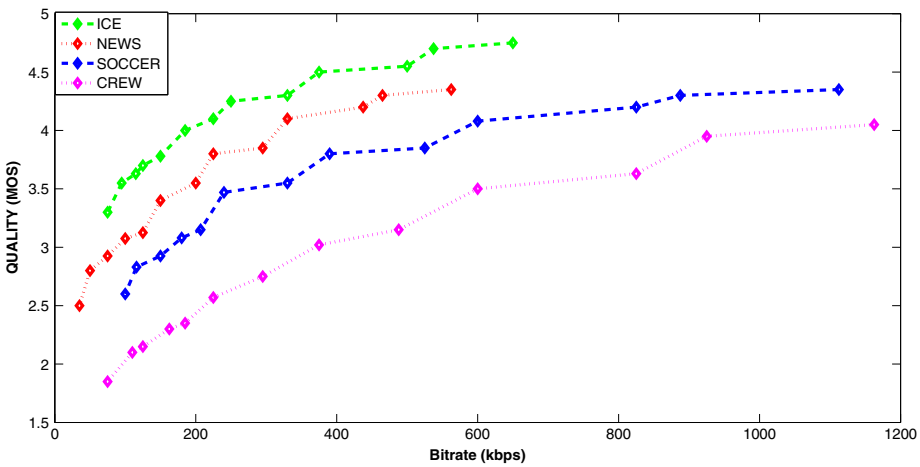

Fig. 9 MOS vs. bitrate characteristics for each of the considered videos sequences with the addition of each SVC quality layer

Table 1 Simulation parameters—downlink LTE scheduling for multi-class traffic

Parameters	Value
Bandwidth, carrier frequency	3 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)
Video resolution	CIF (352 · 288)
Video frame rate	30 FPS
Encoder	JSVM (9.15)

considering a 3-MHz bandwidth). Our main goal is to analyze the performance of the proposed strategy under congestion. Therefore, we simulate a loaded network with 4 *Ice*, 4 *News*, 4 *Soccer*, and 4 *Crew* video streaming users corresponding to an input average traffic rate of 14 Mbps (4.66 bit/s/Hz). In addition, we simulate 4 high-priority video conferencing users, each having average input traffic rate of 200 kbps. The combined input traffic rate is 14.8 Mbps (4.9333 bit/s/Hz) against the input capacity of 2.266 bit/s/Hz. Thus, the average input traffic rate is approximately more than twice the system capacity. The packet delivery target delay of interactive video flows is 150 ms, whereas for video streaming flows the target delay is 400 ms. These QoS parameters are selected according to the LTE QCI [48]. We propose to use the following strategies:

- Simplified fine granularity (SFG) [13] scheduling based on packet's contribution towards video quality. This strategy is similar to the ones proposed in [9–12]. The algorithm comprises a two step process. At each scheduling epoch, the scheduler sorts packets of a flow based on the ratio of packet's contribution towards video quality and its size. The priority of a flow on each PRB is computed by the product of channel quality and the ratio computed in the previous step. PRB is allocated to the flow maximizing the priority function. A detailed implementation of the SFG scheduling rule is given in [13].
- Proxy-based rate adaptation [50]: We assume that a proxy is located close to the eNodeB. The proxy responds quickly to the dynamic wireless channel and congestion by performing rate adaptation. In order to perform rate adaptation, the proxy considers the rate-quality trade-off model, Fig. 9, of the video streaming sequences, the channel quality, and the buffer status of all the video streaming flows. The main goal of the proxy is to maximize the sum MOS associated with different SVC streams based on the periodic congestion signal from the eNodeB. The eNodeB utilizes a QoS-aware M-LWDF packet scheduler at the eNodeB which considers the target packet delivery needs of different traffic types. A detailed implementation of the M-LWDF is given in [39].
- Proposed framework: QoE-based packet marking is performed at the core network by utilizing the content-dependent utility functions (quality vs. bitrate function for each of the considered video sequences) shown in Fig. 9. The mapping of SVC layers into priority classes is shown in Fig. 10. The total number of priority classes is 13. According to the figure, 13 video layers (1 base and 12 quality layers) are mapped into 12 priority classes. For

Fig. 10 QoE-aware priority layer marking for 16 H.264/SVC video streams. SVC layers are marked based on their contribution to the overall QoE according to the greedy search algorithm in [34]

instance if we consider Ice 1 video flow, the base layer is assigned priority class 12 and the first quality sub-layer (SVC layer 2) is assigned priority class 11. Similarly, the last 3 quality layers (SVC layers 11, 12, and 13) are assigned priority class 1. The video conferencing flows are assigned the most important priority class, i.e., class index 13 whereas priority classes 1 to 12 are assigned to the SVC layers according to the mapping shown in Fig. 10.
Rate adaptation is performed via the joint operation of scheduler and priority class filter. Hysteresis-based priority class filtering decisions are taken every 250 TTIs ($t_w = 250$ ms, 4 rate adaptation decision points per second). Furthermore, the higher limit of the threshold S_{thr_h} is set to 0.6. Once the average normalized system delay crosses 0.6 (averaged over 250 TTIs), the probability of delay bound violation $P(S_{\text{thr}_\text{h}})$ is maximum, i.e., 1. On the other hand, the lower limit S_{thr_l} is set to 0.2 (averaged over 250 TTIs) with lower limit probability of delay bound violation $P(S_{\text{thr}_\text{l}})$ set to 0.1. Re-admission speed is set to 0.02, i.e., $\omega = 0.02$.

- QoE-aware packet dropping [18]: In this strategy, QoE-aware packet marking is performed at the P-GW. The marking information is utilized at the eNodeB by dropping low-priority packets under the event of eNodeB congestion. The packet dropping algorithm drops the QoE-marked priority packets before they are fed into the scheduler. The function of the packet dropping algorithm is to shape the traffic according to the wireless transmission capacity, whereas the function of the scheduler is to perform packet scheduling onto the radio resources. It is important to note the scheduler is not aware of the QoE marking. Therefore, we utilize a packet delay-aware M-LWDF scheduling rule.
- No rate adaptation: We assume that there is no admission control policy and the system runs under overloaded condition. We selected a link-efficient Best-CQI scheduler which assigns PRBs based on the channel quality. In addition to the rate maximizing scheduling strategy, we also consider a QoS-based M-LWDF scheduling strategy. According to [51], M-LWDF is the best scheduling rule for delay-sensitive applications in terms of fairness and efficiency.

7.2 Simulation scenario 2

In order to evaluate the QoS/QoE performance of the delay-sensitive traffic under the presence of best-effort traffic, 4 full buffer best-effort flows are added to the network in addition to the 16 video streaming and 4 video conferencing flows of the previous scenario. A full buffer source is greedy in nature having an infinite number of packets in the buffer. Therefore, we aim to analyze the throughput performance of the best-effort traffic under the presence of bandwidth-intensive and delay-constrained video applications by using the proposed composite scheduling rule in Section 6. We compare the proposed rule with the logarithmic and the exponential scheduling rules, proposed in [44], at the eNodeB with proxy-based rate adaptation. A detailed working of these multli-class scheduling algorithms is given in [39, 44]. These rules schedule the best-effort traffic by using the classical proportional fair rule and prioritize the delay-sensitive traffic by the logarithmic (LOG-RULE) and the exponential (EXP-RULE) function of the HoL delay.

The simulation results of the aforementioned strategies are analyzed in the following section.

8 Results

The subsequent sections report the performance of the proposed scheme in comparison to the state-of-the-art approaches, in different scenarios.

8.1 Joint scheduling and priority class filter performance under scenario 1

Figure 11 shows the average system delay when all the video streaming and video conferencing flows enter the system and the wireless access system becomes highly congested. The figure also shows the percentage of flows blocked for the priority classes over the high load simulation period. According to the figure, there is a significant increase in the system load after 5 s. Under congestion, the normalized average system delay decreases the resource allocation probability of the least important priority class (priority class 1). This leads to an increase in the delay bound violations of the least important priority class. The normalized system delay along with the packet loss ratio is calculated over the moving average window. These two parameters are utilized in the priority class blocking of the video flows.

After 5.4 s, the normalized averaged system delay increases above the upper threshold of the hysteresis window with very high system PLR in priority class 1 which triggers the blocking of the flows. After blocking priority class 1, the reduction in the normalized system delay is minimal which does not decrease the delay bound violation of the current least important priority class. According to the figure, all the flows of priority class 2 are blocked in the subsequent priority class filter cycles. After blocking of priority classes 1 and 2, there is no reduction in the normalized system delay and PLR which is an indication that further rate adaptation is required. In the next filter cycle, 55.55 % of the flows of priority class 3 are blocked. According to Fig. 10, priority class 3 has 9 video flows. Therefore, 5 of the 9 video flows are blocked for this class. The priority class filter exploits the proportional

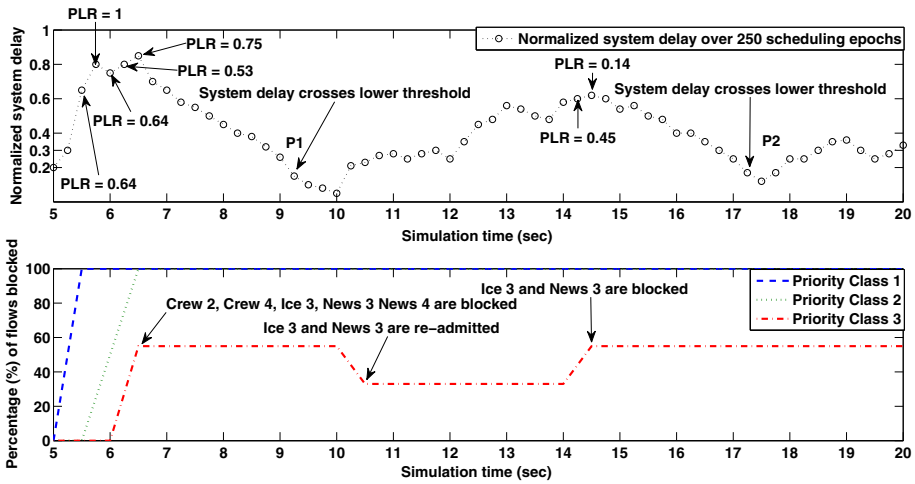

Fig. 11 Priority class blocking based on the hysteresis-based filter algorithm

fair rule and blocks 5 flows having lower channel quality and higher throughput as shown in Fig. 11. After blocking these flows, the average system delay remains between the two thresholds till point "P1" (9.25 s). This is an indication that the input arrival rate is within the achievable rate region according to the delay bound constraints.

The average system delay crosses the lower threshold at 9.25 s. The hysteresis output decreases exponentially, according to (15), with each window cycle having system delay below the lower threshold. The lower the hysteresis output, the higher the re-admission probability. The filter re-admits the ICE 3 and News 3 flows which increases the average system delay, thus inhibiting the re-admission of further flows for priority class 3. It is important to note at point "P2" that the system delay crosses the lower threshold for approximately 0.5 s. The decrease in the hysteresis output is not sufficient enough to re-admit flows. The hysteresis-based process adds stability and decreases the variations in perceivable video quality due to variations in the input arrival traffic and wireless link capacity.

The achieved MOS and the associated channel quality for each of the video users are shown in Fig. 12. According to the figure, the proxy-based strategy performs best, in terms of the achieved MOS, followed by the

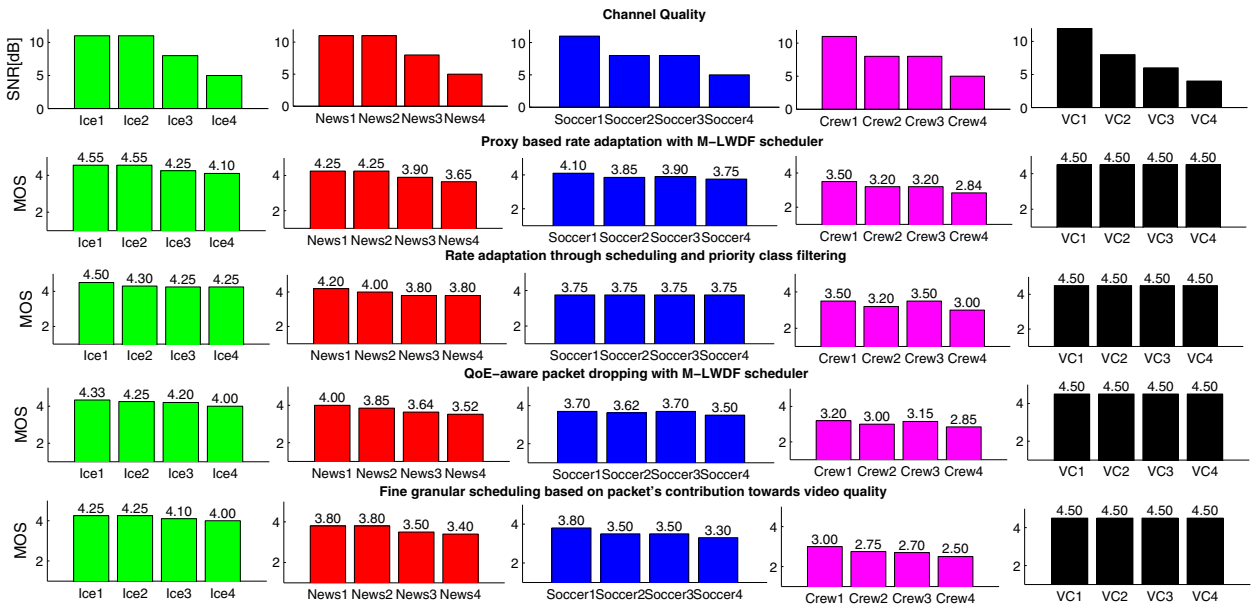

Fig. 12 MOS performance for each of the video users under scenario 1

proposed joint scheduling and filter policy relying on QoE marking.

It is important to note that the system schedules priority classes 4 to 13 of all the flows with zero PLR. The PLR, due to scheduling (delay bound violation) and priority class filtering, of SVC layers is shown in Fig. 13. According to the figure, all the *soccer* flows receive the same number of SVC layers. This is mainly due to the fact that the filter blocks priority classes 11 and 12 for all the video flows as shown in Fig. 11. Therefore, 5 enhancement layers (SVC *IDs* 9 to 13) are blocked by the filter for all the *soccer* flows. If we analyze all the *news* flows, the SVC layers assigned to priority class 3 for *news* 1 and *news* 2 flows are never blocked by the filter because of their good channel quality. Similar analysis holds for the *ice* 1 and *ice* 2 flows. These 4 flows' priority class 3 remains unblocked by the filter throughout the simulation period. At the 14.5-s simulation time, there is a PLR of 0.14. However, system delay and the PLR are not high enough to block any further flows for priority class 3. Therefore, a PLR lower than 3 % appears for the SVC layers 9, 10, 11, and 12 (assigned to priority class 3) for *news* 1 and *news* 2 flows.

8.1.1 Performance comparison with the proxy based strategy

Figure 12 also reports the MOS performance of the proxy-based rate adaptation scheme with M-LWDF scheduler at the eNodeB. According to Fig. 12, most of the video streaming users achieve better MOS as compared to the proposed strategy. This is mainly due to the fact that the proxy never overloads the scheduler by considering the channel quality and the throughput requirements of each flow. Therefore, the input arrival rate at the eNodeB is always within the achievable rate region. When the input traffic goes above the input arrival rate, the proxy drops the video layers contributing lowest to the video quality for the flows having poor channel quality.

The proxy-based solution requires explicit signaling, every second, in order to collect the channel quality of all the video flows from the eNodeB. However, the channel quality of the mobile users can change significantly over the period of 1 s. Therefore, the proxy can receive outdated channel quality of the mobile users, which can limit the video streaming performance. For instance, the average per user MOS (sum MOS, Fig. 12, of all the video flows divided by the total number of video flows) for the QoE marking-based rate adaptation is approximately 3.965 compared to 3.992 for the proxy-based rate adaptation. The difference in the average MOS is only 0.027. The smaller performance difference mainly stems from the fact that proxy-based rate adaptation has slower congestion avoidance frequency in response to the stochastic nature of the wireless channel. On the other hand, the joint scheduling and filter policy regulates the system load at a much faster interval (after every t_w scheduling epochs). Therefore, the proposed strategy has a quick rate adaptation response which suits the highly dynamic mobile environment. The main reason of a slight superior performance of the proxy-based scheme is its proactive nature, i.e., it allows the scheduler to run optimally by keeping the input traffic according to the system capacity. On the other hand, the proposed joint strategy is reactive in nature, i.e., the rate adaptation is performed when a delay bound violation occurs. However, because of the quick rate adaptation response (4 cycles per seconds) the performance penalty in terms of average MOS is minimal, i.e., 0.027.

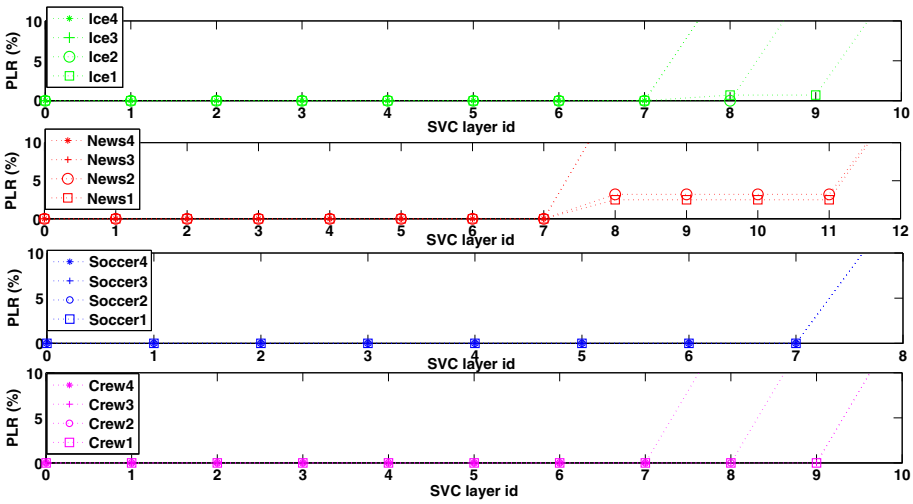

Fig. 13 PLR (%) on each of the SVC layers with the proposed joint scheduling and filter policy. Where the PLR value for a layer is not shown in the figure, the relevant layers have been blocked by the filter

8.1.2 Performance comparison with the QoE-aware packet dropping strategy

The QoE-aware packet dropping strategy results in per user average MOS of 3.8255 as compared to 3.965 for the proposed rule. The dropping algorithm in [18] computes average achievable bitrate of each user by considering the downlink channel quality. Each priority class of the video flows is assigned a transmission score. The transmission score is the product of downlink channel quality and QoE contribution of the priority class. According to the average achievable bitrate, priority classes of each video flow are transferred to the scheduler in the order of their transmission score. The remaining priority classes are dropped. The algorithm works at a time scale of 1–2 s. In mobile environment, channel quality can change considerably over the time scale of 1–2 s. Therefore, underutilization of radio resources occurs when the channel quality of mobile users is improved after the packet dropping cycle. Similarly, packet delay bound violation occurs when the channel quality of mobile users fades after the packet dropping cycle. This can result in delay bound violation of high-priority packets since the scheduler is unaware of the priority class importance. On the other hand, the proposed PPS scheduler prioritizes higher priority classes under congestion. The proposed rule is robust to the fluctuations in wireless channel quality by employing a fast rate adaptation cycle. In addition, the hysteresis principle adds stability in the rate adaptation decisions.

8.1.3 Performance comparison with the SFG rule

The average per user MOS for the SFG scheduling rule is 3.7075 compared to 3.965 for the proposed strategy. The significant performance difference between the two strategies is mainly due the fact that the SFG scheduling rule is packet delay agnostic. The SFG rule is unable to determine the scheduling urgency of packets nearing the maximum tolerable delay bound. Another important point to consider is that SFG rule has no prioritization framework between interactive and streaming video traffic. In order to accommodate interactive video, SFG uses the strict prioritization scheduling rule which reduces the channel diversity exploitation between the flows of the two traffic classes. On the other hand, the PPS scheduling function becomes strictly priority aware only under higher normalized system delay. Under congestion, the priority class filter reduces the normalized system delay by blocking least important priority classes which increases the channel awareness of the scheduling rule, thus increasing the channel diversity exploitation between the flows of two traffic classes.

8.1.4 Performance comparison with the QoE-unaware rules

For the considered scenario, the performance of the M-LWDF and the Best-CQI scheduling rule with no rate adaptation scheme is shown in Fig. 14. When the system is left to run under high load, the M-LWDF scheduling rule only serves higher priority flows (video conferencing users) with good quality. Therefore, all the delay-aware scheduling rules must ensure that the arrival rate should not exceed the system capacity, otherwise the QoS performance of the existing users in the network would be violated resulting in an increase in the number of unsatisfied users. In the considered scenario, the average input arrival rate is 14.8 Mbps as compared to the

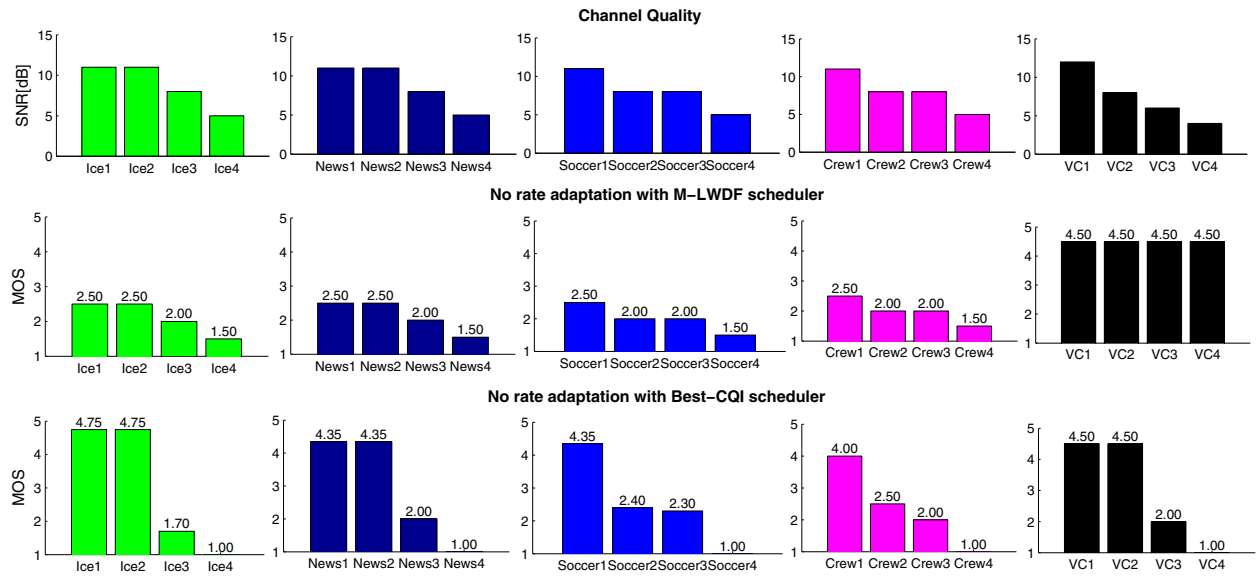

Fig. 14 MOS performance for each of the video users under scenario 1 (no rate adaptation strategies are used as the benchmark)

system capacity of 7 Mbps. There is no rate adaptation policy for the M-LWDF scheduling rule, therefore this scheduling rule requires a proper flow admission control policy which should not increase the arrival rate above the system capacity. The increase in arrival rate above the system capacity incurs delay bound violations. Furthermore, the least important packets reside in the buffer till the delay bound, block packets of important priority classes. This head-of-line blocking of important priority classes reduces the QoE of all the video streaming flows. Therefore once the delay-sensitive traffic's arrival rate reaches the system capacity, the admission control policy should block further flows from entering the system. Similar analysis holds for the Best-CQI scheduler. However, the Best-CQI scheduler favors the good channel quality flows and results in an unfair system as shown in Fig. 14.

8.2 Performance of the proposed and benchmark strategies under scenario 2

Figure 15 shows the normalized system delay, sum MOS, and sum throughput performances of the proposed composite scheduling rule and the benchmark strategies. The average per user MOS of the proposed rule, without best-effort flows, is 3.965 for 20 delay-sensitive flows as reported in the previous scenario. This decreases to 3.945 (sum MOS = 78.9 with C = 50) and 3.917 (sum MOS = 78.35 with C = 75 and 100) as shown in Fig. 15. The decrease in the average MOS with and without best-effort flows is minimal. It is important to note that at lower system delays, the priority weight in (20) increases the resource allocation probability of the best-effort flows. Therefore, the normalized system delay of delay-sensitive flows crosses below the lower threshold for a shorter period of time as shown in Fig. 15. This decreases the probability of re-admission of blocked least important priority classes as the re-admission speed depends upon the normalized system delay of delay-sensitive flows. The lower the normalized system delay, the higher the probability of re-admission of blocked classes of the delay-sensitive flows.

According to the QoE marking for SVC streaming flows, the least important priority classes contribute less towards the overall QoE. Thus, the impact of the best-effort flows on the QoE of the delay-sensitive flows is negligible. When the prioritization factor increases from 50 to 75, the amount of time normalized system delay remains below the lower threshold is decreased further. The sum MOS of delay-sensitive flows decreases minimally from 78.9 to 78.35 with the increase in sum throughput to 682.45 kbps. A further increase in the prioritization factor, 75 to 100, has no impact on the performance of the delay-sensitive and best-effort traffic. This shows that the system delay remains between the two thresholds, with

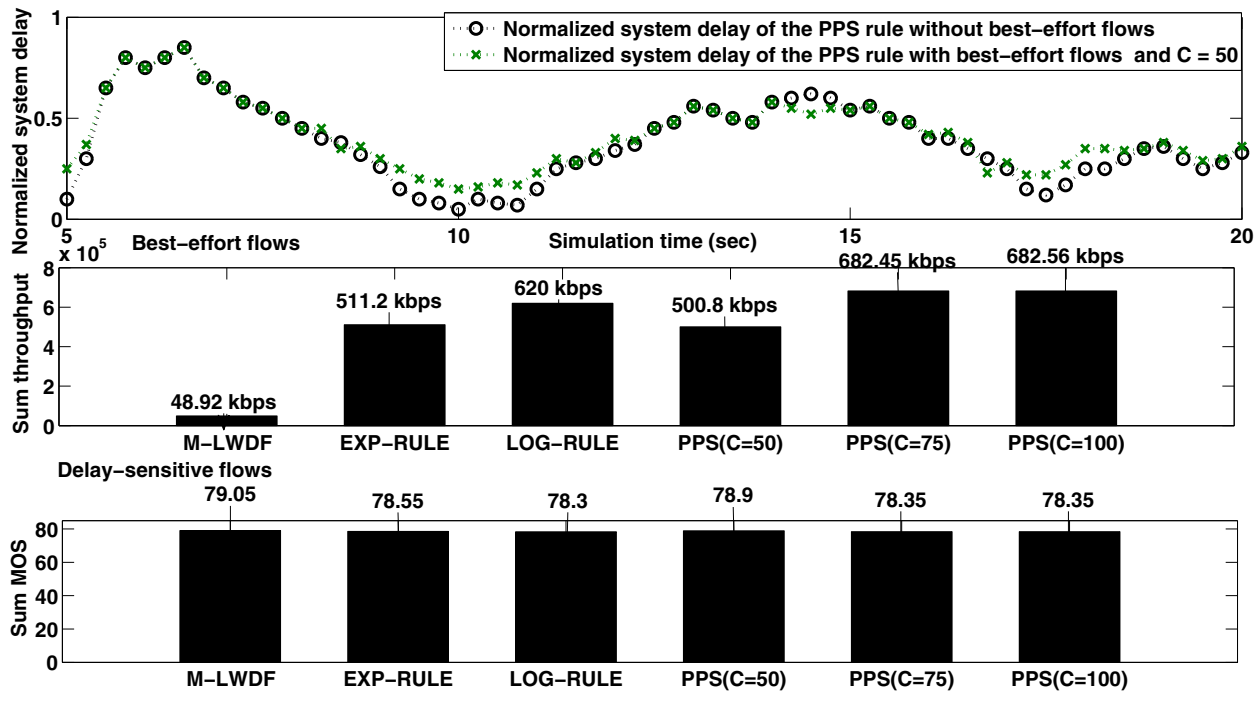

Fig. 15 *Above*: normalized system delay performance with and without best-effort flows; *below*: sum throughput and sum MOS performance for best-effort and delay-sensitive flows (under scenario 2), respectively

delay-sensitive flows meeting their desired QoS/QoE and best-effort flows maximizing the system throughput.

The figure also reports the performance of the state-of-the-art composite scheduling rules. M-LWDF rules achieve higher prioritization for the delay-sensitive flows with best-effort flows receiving no service. On the other hand, the log and exponential rules achieve better inter-class fairness. Exponential and the logarithmic scheduling rules reduce the starvation of the best-effort traffic which causes an increase in the buffer level of the delay sensitive flows. The increased queue status of the delay-sensitive flows is periodically monitored by the proxy, resulting in rate adaptation for the SVC flows at the proxy, which causes a decrease in the sum MOS for delay-sensitive flows.

9 Conclusions

Network operators are faced with rapidly growing video traffic that is becoming a main source of congestion in their networks. In order to reduce congestion, timely video rate adaptation is required at the RAN. This requires substantial investments on new modules which can reduce the congestion through fast rate adaptation of video traffic at the RAN. In this work, we propose a novel PPS, aiming at minimizing delay bound violations for the most important priority classes, and a priority class filter strategy, operating jointly with the scheduler to provide video rate adaptation at the MAC layer. The priority class filter utilizes parameters of the scheduling function and provides fast video rate adaptation at the MAC layer, without requiring additional modules at the RAN. The proposed framework is capable of reducing congestion and provide high QoE for delay-sensitive as well as the delay-tolerant traffic classes. Simulation results show that operators can provide guaranteed services to high-priority delay-sensitive flows marked with the highest priority class index. Furthermore, the priority class filtration of QoE-based SVC classes reduces the resource starvation of best-effort traffic and ensures that best-effort flows maximize the system throughput subject to the QoS/QoE constraints of delay-sensitive flows.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7) under grant agreement 288502 (CONCERTO). The authors would like to acknowledge Prof. Dirk Staehle for the valuable discussions.

Received: 15 August 2015 Accepted: 10 March 2016
Published online: 01 April 2016

References

1. Cisco Systems, Cisco Visual Networking Index: global mobile data traffic forecast update, 2014-2019. White Paper (2015)
2. H Schwarz, D Marpe, T Wiegand, Overview of the scalable video coding extension of the H.264/AVC standard. IEEE Trans. Circ. Syst. Video Technol. **17**(9), 1103–1120 (2007)
3. GJ Sullivan, JM Boyce, Y Chen, JR Ohm, A Segall, A Vetro, Standardized extensions of High Efficiency Video Coding (HEVC). IEEE J. Sel. Top. Sign. Process. **7**(6), 1001–1016 (2013)
4. ETSI TR 102 643 V101, Human factors (HF); quality of experience (QoE) requirements for real-time communication services. Tech. rep., European Telecommunications Standards Institute (2009)
5. ITU-R BT 500-11, Recommendation, methodology for the subjective assessment of the quality of television pictures. Tech. rep., International Telecommunication Union (2002)
6. ZN Li, MS Drew, J Liu, *Fundamentals of multimedia*. (Springer, 2014)
7. Z Wang, AC Bovik, HR Sheikh, EP Simoncelli, Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
8. M Razaak, M Martini, K Savino, A study on quality assessment for medical ultrasound video compressed via HEVC. IEEE J. Biomed. Health Inform. **18**(5), 1552–1559 (2014)
9. F Li, D Zhang, M Wang, Multiuser multimedia communication over orthogonal frequency-division multiple access downlink systems. Concurr. Comput. Pract. Experience. **25**(9), 1081–1090 (2013)
10. F Li, G Liu, L He, A low complexity algorithm of packet scheduling and resource allocation for wireless VoD systems. IEEE Trans. Consum. Electron. **56**(2), 1057–1062 (2010)
11. P Li, Y Chang, N Feng, F Yang, A cross-layer algorithm of packet scheduling and resource allocation for multi-user wireless video transmission. IEEE Trans. Consum. Electron. **57**(3), 1128–1134 (2011)
12. F Li, P Ren, Q Du, Joint packet scheduling and subcarrier assignment for video communications over downlink OFDMA systems. IEEE Trans. Veh. Technol. **61**(6), 2753–2767 (2012)
13. Y Zhang, G Liu, Fine granularity resource allocation algorithm for video transmission in orthogonal frequency division multiple access system. IET Commun. **7**(13), 1383–1393 (2013)
14. S Thakolsri, W Kellerer, S Khan, E Steinbach, QoE-driven cross-layer optimization for high speed downlink packet access. J. Commun. **4**(9), 669–680 (2009)
15. S Thakolsri, W Kellerer, E Steinbach, in *International Conference on Communications and Networking in China (ChinaCOM)*. Application-driven cross layer optimization for wireless networks using MOS-based utility functions, (Xi An, 2009)
16. S Cicalò, V Tralli, Distortion-fair cross-layer resource allocation for scalable video transmission in OFDMA wireless networks. IEEE Trans. Multimed. **16**(3), 848–863 (2014)
17. A Ahmedin, K Pandit, D Ghosal, A Ghosh, in *International Conference on Distributed Computing and Networking (ICDCN)*. Exploiting scalable video coding for content-aware downlink video delivery over LTE (Springer, Coimbatore, 2014), pp. 423–437
18. B Fu, D Staehle, G Kunzmann, E Steinbach, W Keller, QoE-based SVC layer dropping in LTE networks using content-aware layer priorities. ACM Trans. Multimed. Comput. Commun. Appl. **18**(5), 1–23 (2014)
19. N Khan, M Martini, in *IEEE International Conference on Communications (ICC) - Workshop on Smart Communication Protocols and Algorithms*. Hysteresis based rate adaptation for scalable video traffic over an LTE downlink, (London, 2015)
20. N Khan, MM Nasralla, M Martini, in *IEEE International Conference on Communications (ICC) - Workshop on Quality of Experience-based Management for Future Internet Applications and Services (QoE-FI)*. Network and user centric performance analysis of scheduling strategies for video streaming over LTE, (London, 2015)
21. G Liebl, T Stockhammer, C Buchner, A Klein, in *Packet Video Workshop*. Radio link buffer management and scheduling for video streaming over wireless shared channels, (Irvine, 2004)
22. G Liebl, H Jenkac, T Stockhammer, C Buchner, Radio link buffer management and scheduling for wireless video streaming. Telecommun. Syst. Springer Sci. Bus. Media B.V. **30/1–3**, 255–277 (2005)
23. PV Pahalawatta, R Berry, TN Pappas, AK Katsaggelos, Content-aware resource allocation and packet scheduling for video transmission over wireless networks. IEEE J. Sel. Areas Commun. (JSAC). **25**(4), 749–759 (2007)
24. L Choi, W Kellerer, E Steinbach, On cross-layer design for streaming video delivery in multiuser wireless environments. Eurasip J. Wirel. Commun. Netw. **2006**, 060349 (2006)

25. N Khan, MG Martini, Z Bharucha, in *IEEE International Workshop on Signal Processing and Advances in Wireless Communications (SPAWC)*. Quality-aware fair downlink scheduling for scalable video transmission over LTE systems, (Cesme, 2012)
26. N Khan, MG Martini, D Staehle, in *IEEE Vehicular Technology Conference (VTC)*. Opportunistic proportional fair downlink scheduling for scalable video transmission over LTE systems, (Las Vegas, 2013)
27. Y Ju, Z Lu, D Ling, X Wen, W Zheng, W Ma, QoE-based cross-layer design for video applications over LTE. Multimed. Tools Appl. **72**(2), 1093–1113 (2014)
28. M Rugelj, U Sedlar, M Volk, J Sterle, M Hajdinjak, A Kos, Novel cross-layer QoE-aware radio resource allocation algorithms in multiuser OFDMA systems. IEEE Trans. Commun. **62**(9), 3196–3208 (2014)
29. ITU-R BT 500-13, Recommendation, methodology for the subjective assessment of the quality of television pictures. Tech. rep., International Telecommunication Union (2012)
30. ITU-T P910, Recommendation, Subjective video quality assessment methods for multimedia applications. Tech. rep., International Telecommunication Union (2008)
31. MH Pinson, S Wolf, A new standardized method for objectively measuring video quality. IEEE Trans. Broadcast. **50**(3), 312–322 (2004)
32. P Amon, T Rathgen, D Singer, File format for scalable video coding. IEEE Trans. Circ. Syst. Video Technol. **17**(9), 1174–1185 (2007)
33. S Wenger, Y Wang, T Schierl, Transport and signaling of SVC in IP networks. IEEE Trans. Circ. Syst. Video Technol. **17**(9), 1164–1173 (2007)
34. B Fu, D Staehle, G Kunzmann, E Steinbach, W Kellerer, in *8th ACM workshop on performance monitoring and measurement of heterogeneous wireless and wired networks*. QoE-aware priority marking and traffic management for H.264/SVC-based mobile video delivery, (Barcelona, 2013)
35. GJ Sullivan, JR Ohm, WJ Han, T Wiegand, Overview of the high efficiency video coding standard. IEEE Trans. Circ. Syst. Video Technol. **22**(12), 1648–1667 (2012)
36. S Tsai, A Soong, Effective-SNR mapping for modeling frame error rates in multiple-state channels. 3GPP2, Tech. Rep. 3GPP2-C30-20030429-010 (2003)
37. C Mehlfuhrer, M Wrulich, JC Ikuno, D Bosanska, M Rupp, in *European Signal Processing Conference (EUSIPCO)*. Simulating the long term evolution physical layer, (Glasgow, 2009)
38. C Mehlfuhrer, JC Ikuno, M Simko, S Schwarz, M Wrulich, M Rupp, The Vienna LTE simulators—enabling reproducibility in wireless communications research. EURASIP J. Adv. Sig. Process. **2011**, 29 (2011)
39. 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. Tutor. **15**(2), 1–23 (2012)
40. CY Wong, RS Cheng, KB Letaief, RD Murch, Multicarrier OFDM with adaptive subcarrier, bit, and power allocation. IEEE J. Sel. Areas Commun. **17**(10), 1747–1758 (1999)
41. I Kim, HL Lee, B Kim, YH Lee, in *IEEE Global Communications Conference (GLOBECOM)*. On the use of linear programming for dynamic subchannel and bit allocation in multiuser OFDM, (San Antonio, 2001)
42. M Andrews, K Kumaran, K Ramanan, A Stolyar, P Whiting, R Vijayakumar, Providing quality of service (QoS) over a shared wireless link. IEEE Commun. Mag. **39**(2), 150–154 (2001)
43. 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)
44. B Sadiq, R Madan, A Sampath, Downlink scheduling for multiclass traffic in LTE. EURASIP J. Wirel. Commun. Netw. **2009**, 510617 (2009)
45. N Khan, MG Martini, D Staehle, QoS-aware composite scheduling using fuzzy proactive and reactive controllers. EURASIP J. Wirel. Commun. Netw. **2014**, 138 (2014)
46. N Khan, MG Martini, D Staehle, in *IEEE Vehicular Technology Conference (VTC)*. QoS-aware fair downlink scheduling for delay sensitive applications using fuzzy reactive and proactive controllers, (Las Vegas, 2013)
47. N Khan, MG Martini, Z Bharucha, G Auer, in *IEEE Wireless Communications and Networking Conference (WCNC)*. Opportunistic packet loss fair scheduling for delay-sensitive applications over LTE systems, (Paris, 2012)
48. 3GPP TS 23203, Release 9, Group services and system aspects—policy and charging control architecture. Tech. rep., 3GPP (2009)
49. C Horch, J Habigt, C Keimel, K Diepold, in *IEEE International Workshop on Quality of Multimedia Experience (QoMEX)*. Evaluation of video quality fluctuations using pattern categorization, (Singapore, 2014)
50. H Hu, X Zhu, Y Wang, R Pan, J Zhu, F Bonomi, in *IEEE International Conference on Communications (ICC)*. QoE-based multi-stream scalable video adaptation over wireless networks with proxy, (Ottawa, 2012)
51. HAM Ramli, R Basukala, K Sandrasegaran, R Patachaianand, in *IEEE Malaysia International Conference on Communications (MICC)*. Performance of well known packet scheduling algorithms in downlink 3GPP LTE system, (Kuala Lumpur, 2009)