An intelligent fuzzy logic-based content and channel aware downlink scheduler for scalable video over OFDMA wireless systems

MDPI

Article

An Intelligent Fuzzy Logic-based Content and Channel Aware Downlink Scheduler for Scalable Video over OFDMA Wireless Systems

Peter E. Omiyi [1], Moustafa M. Nasralla [2*], Ikram Ur Rehman [3], Nabeel Khan [4] and Maria G. Martini [5]

[1] Faculty of Engineering, Holon Institute of Technology
[2] Department of Communications and Networks Engineering, Prince Sultan University, Riyadh, Saudi Arabia;
[3] School of Computing and Engineering, University of West London, UK;
[4,5] Faculty of Science, Engineering and Computing, Kingston University;
* Correspondence: mnasralla@psu.edu.sa

Academic Editor: name
Version June 22, 2020 submitted to Electronics; Typeset by LaTeX using class file mdpi.cls

Abstract: The recent advancements of wireless technology and applications make downlink scheduling and resource allocations an important research topic. In this paper, we consider the problem of downlink scheduling for multi-user scalable video streaming over Orthogonal Frequency Division Multiple Access (OFDMA) channels. The video streams are precoded using an scalable video coding (SVC) scheme. We propose a Fuzzy logic-based scheduling algorithm, which prioritizes the transmission to different users by considering video content, and channel conditions. Furthermore, a novel analytical model and a new performance metric have been developed for the performance analysis of the proposed scheduling algorithm. The obtained results show that the proposed algorithm outperforms the content-blind/channel aware scheduling algorithms with a gain of as much as 19% in terms of the number of supported users. The proposed algorithm allows for a fairer allocation of resources among users across the entire sector coverage, allowing for the enhancement of video quality at edges of the cell while minimizing the degradation of users closer to the base station.

Keywords: Content-Aware, Cross-Layer, Fuzzy Inference System, OFDMA, Scheduling, SVC, Video Streaming

1. Introduction

Supporting multimedia applications and services over wireless networks is challenging due to constraints and heterogeneities such as limited bandwidth, limited battery power, random time-varying channel conditions, different protocols and standards, and varying quality of service (QoS) requirements. Two main classifications can be performed as far as scheduling algorithms are concerned: channel-aware schedulers and content-aware schedulers. A comprehensive survey on downlink channel-aware and content-aware scheduling algorithms can be found in [1,2].

It is worth mentioning that channel-unaware schedulers make no use of channel state conditions such as power level, channel error, and loss rates. These basically focus on fulfilling delay and throughput constraints. Examples of the traditional channel-unaware schedulers are Round-Robin, weighted fair queuing (WFQ), and priority-based algorithms. Such algorithms assume perfect channel conditions, no loss, and unlimited power source. However, due to the nature of wireless

medium and the user mobility, these assumptions are not valid. The Base Station (BS) downlink scheduler could rather use channel information (e.g. channel state information (CSI), including the Carrier to Interference and Noise Ratio (CINR)) which is reported back from the mobile receiver. Most of the channel-aware algorithms assume that channel conditions do not change within the frame period. It is also assumed that the channel information is known at both the transmitter and the receiver. In general, schedulers favour the users with better channel quality to exploit the multi-user diversity and channel fading. However, to meet fairness requirements, the scheduler also needs to consider other users' requirements and should introduce some compensation mechanisms.

In content-unaware scheduling strategies, the QoS of the received video is measured in generic terms of packet delay, packet loss rate or data rate. In general, these methods exploit the variability of the wireless channel over time and across users, allocating a majority of the available resources to users with good channel quality. Ultimately, these scheduling strategies support higher data rates, while maintaining fairness across multiple users. In this context, these strategies attempt to maximize a utility function, which is defined as either a function of each user's current average throughput, or of each user's queue length or delay of the head-of-line packet [2].

In contrast, content-aware scheduling strategy is not a simple function of the data rate, delay or data loss but it is rather affected differently by the impact of losses and errors in different segments of the video stream. This is highlighted in an SVC bitstream, which consists of one base layer and multiple enhancement layers. As long as the base layer is received, the receiver can decode the video stream. As more enhancement layers are received, the decoded video quality is improved. In multi-user video transmission, this introduces a type of multi-user content diversity that can be exploited by content-aware scheduling policies in optimizing the utilization of the network resources. Examples of content-aware methods and current SVC studies are found in [2–9].

Unlike state-of-the-art content aware strategies, the proposed scheduling rule considers SVC layer priority index. The higher the layer priority, the higher the probability of the layer to be scheduled. The layers can be marked outside the eNodeB, for instance at the P-GW or video server, whereas the scheduler at the eNodeB exploits the layer priority marking and schedule layers contributing maximum to the overall video quality. There exist several QoE layer marking strategies, such as in [10,11], where SVC layers are marked based on their contribution to the overall QoE. Therefore, the proposed scheduling rule requires only layer priority index at the eNodeB, whereas the complex processing of SVC layer marking is performed outside the eNodeB. On the other hand, the state-of-the-art content aware scheduling requires complex video content processing at the eNodeB. However, the transfer of video content related information to the eNodeB is not practical thus restricting the usage of such strategies.

Several content aware scheduling strategies [12–17] evaluate the value of the content and maximize the video quality of the streaming users subject to the channel constraints. However, such strategies suffer from high computation complexity at the MAC layer of the eNodeB. In order to address the complexity issue, we proposed a scheduling strategy, where complexity in terms of the number of iterations varies linearly *w.r.t* the number of users and resources. In other words, the proposed fuzzy-based scheduling priority function is a linear function of the users (competing for resources) and the number of resources. This is in contrast to the content-aware scheduling strategies where scheduling complexity varies exponentially *w.r.t* the number of users and resources.

Furthermore, the literature lacks proposals on content-aware priority-based scheduling algorithm which utilizes an Fuzzy Inference System (FIS). An FIS considers the concept of vagueness and uses probability-based mathematical models to represent the vagueness. Words/estimates are potentially less precise than numbers or Boolean representation; however, words are closer to human intuition. Hence, FIS would be a good approach to explore the tolerance for imprecisions and hence gain a better understanding of the application. There are two common inference methods: the Mamdani's fuzzy inference method and the Takagi-Sugeno-Kang method of fuzzy inference. In several studies related to real time scheduling, as in [18,19], it was proven that Mamdani-type FIS and

Sugeno-type FIS perform similarly, except that using Sugeno-type FIS model allows the scheduling system to work at its full capacity. In addition, it was proven that Sugeno-type FIS has the advantage that it can be integrated with neural networks and genetic algorithm or other optimization techniques, so that the controller can adapt to individual user and variable channel conditions [18,19]. FIS is an effective tool to establish relationships between input and output variables. It is particularly useful for relatively small dataset and limited number of input variables. Utilising FIS, we propose a downlink scheduling algorithm and a user utility function, which complements our study. Furthermore, this method provides computational efficiency and is well-suited for optimization and adaption of algorithms, which makes it a potential candidate for scheduling problems, in particular for dynamic wireless systems. Hence, for our study the popular Sugeno's FIS method is chosen.

This paper provides four main contributions, highlighted below:

1. Proposing a multi-user content-aware priority-based scheduling algorithm, where packet priorities are selected based on Sugeno FIS.
2. Proposing a framework for quantitatively classifying the video content, in order to apply the proposed FIS.
3. Proposing a performance metric called significance throughput. This metric gives a better indication of the scheduler performance for content sensitive traffic than throughput.
4. Lastly, proposing a novel analytical model of the FIS-based scheduling algorithm, and providing analysis of it.

The rest of the paper is organized as follows. Section 2 provides a background on fuzzy inference system, OFDMA systems, and scalable video coding. In Section 3, we present the related work on the existing downlink scheduling algorithms. Section 4 describes the methodology which consists of an FIS-based downlink scheduling algorithm, a wireless system model, a novel key performance metric (i.e. significance throughput), and lastly, the analytical model to analyse the proposed scheduling algorithm. Results of the analytical model are reported in Section 5. Finally, Section 6 concludes the paper.

2. Background

In this section, we provide background on the core aspects of this paper which are fuzzy inference system, OFDMA systems, and scalable video coding.

As mentioned earlier in Section 1, for our study the Sugeno fuzzy inference method is chosen. The underlying concept of FIS is that of a linguistic variable which makes it closer to human intuition. Hence, fuzzy logic is a good approach to explore the tolerance for imprecisions and hence gain a better understanding of the application. An FIS performs the mapping of a given input to an output using the Fuzzy Logic and by employing components such as membership functions, fuzzy logic operators and If-Then rules. After the input and output variables are defined for the Fuzzy system, the next step is to assign linguistic labels in order to provide quantification of the values, which are defined through membership functions. More details on the functionality of FIS can be found in [20] [21].

The underlying wireless technology considered in this paper is 4G system, which is based on OFDMA. The OFDMA systems allow multiple users to share the spectrum at the same time. The subcarriers in OFDMA are shared between multiple users; to enable better utilisation of radio resources. This technique helps wireless technologies improve the system capability to achieve the following: 1) support high data rates, 2) provide multi-user diversity, 3) compact/eliminate the Inter-Symbol-Interference (ISI) caused by multipath fading, and 4) to be immune to frequency selective fading [2].

The video streams used in this paper are precoded using an SVC scheme. SVC [22] represents a video sequence via multiple layers with different quality, resolution, and/or frame rate as shown in Figure 1. SVC enables graceful degradation of video quality when resources are limited, hence it is particularly suitable for the case of multi-user video scheduling. In other words, an SVC stream

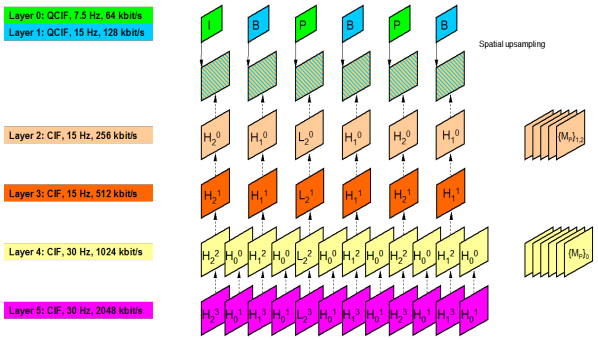

Figure 1. Temporal, Spatial and Quality scalability of SVC.

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, the decoded video quality is improved. The scalability of SVC consists of temporal scalability, spatial scalability, and quality scalability. In this work, we consider SVC with temporal scalability, however our approach is also applicable to other scalability models which are defined in [22]. For example, in temporal scalability model, we consider a Hierarchical B frame GOP structure as follows $\{K_0B_2B_1B_2K_0..\}$, where K_0 is an I or P key picture. The number of coding layers $N_L = 3$. Each layer is composed of one or more frames. The layers in order of importance are the key picture $\{K_0\}$ with index $l = 0$, $\{B_1\}$ with index $l = 1$, and $\{B_2B_2\}$ with index $l = 2$. The significance values are $v = 1, v = 2/3$ and $v = 1/3$, respectively.

3. Related Work

Over the years, various packet scheduling algorithms have been developed to support Real Time (RT) and non-Real Time (NRT) services, comprising the most commonly used ones, namely: Proportional Fair (PF), Modified Largest Weighted Delay First (M-LWDF) and Exponential-PF (EXP-PF) schedulers [2,3]. In the aforementioned schedulers, each flow is assigned a priority value and the radio bearer which carries the flow with the highest priority value will be scheduled first at the corresponding Transmission Time Interval (TTI). When transmitting multimedia services to multiple users over wireless systems, a scheduling strategy should address the trade-off between resource utilization and fairness among users. Network operators are mostly interested in maximizing the exploitation of the resources, e.g., assigning more resources to the user(s) experiencing better channel conditions. However, this approach of theirs can result in unsatisfied users, which in turn would result in users' experiencing worse channel conditions and hence, leading towards not meeting their QoS and Quality of Experience (QoE) requirements.

In our previous studies [2,23], we carried out a comprehensive review on the existing content-aware strategies. In addition, we classified content-aware strategies into the following three classes: 1) Quality driven scheduling approach, 2) Proxy driven radio resource allocation approach, and 3) Client driven approach. In this paper, we take a step forward on proposing a content-aware scheduling strategy that would fall under the first class i.e., Quality driven scheduling approach as this approach consists of scheduling strategies specifically designed for video streaming traffic. In this approach, the information on the content of different video traffic flows is provided through cross-layer signaling to the Radio Access Network (RAN). These types of schedulers consider in their scheduling decision different objective functions (e.g., mean square error (MSE), peak signal-to-noise ratio (PSNR), and Structural Similarity (SSIM)) based on the video quality. The main goal of this scheduler is to maximize the video quality of the streaming users under channel and bandwidth constraints.

Content-aware downlink packet scheduling schemes for multi-user scalable video delivery over wireless networks are proposed in [24–26]. Their schedulers use a gradient-based scheduling

framework along with SVC schemes. Similarly, a content-aware and fair downlink packet scheduling algorithm for scalable video transmission over Long-Term Evolution (LTE) systems is proposed in [27]. The authors proposed a Nash bargaining based on fair downlink scheduling strategy for scalable video transmission to multiple users. A novel utility metric based on the importance of the video contents obtained from a Group of Pictures (GOP) is used in conjunction with the decoding deadline of the GOP. The system capacity in terms of satisfied users can be increased by 20% with the proposed content-based utility in comparison with advanced, state-of-the-art throughput based strategies. The authors in [28] improve the work in [27] by exploiting multi-user time-averaged diversity. The reason for using SVC is to provide multiple high quality video streams over different prevailing channel conditions for multiple users. The schedulers proposed outperform the traditional content-blind scheduling approaches. Furthermore, a significant improvement was observed in terms of objective video quality metrics (e.g.Throughput, PSNR and SSIM etc.) when the proposed scheduling schemes were compared with the content-blind scheduling schemes in the presence of network congestion. Hence, it was established that the video content should be given utmost importance after QoS, when determining the quality of video sequences [3]. However, the proposed content-aware schedulers did not explicitly consider channel conditions in its allocation process. In a wireless environment, this could lead to poor video quality, with a few users with very poor channel conditions, using almost all the available channel resources to satisfy their video quality requirements.

It is worth mentioning that video quality is subjective, and while it is relatively straightforward to distinguish between the importance of different segments of the video stream, based on their relative impact on video quality, it is difficult to quantify these differences. In [3,29–31], priority-based scheduling algorithms are proposed, with the priority function taking into account the importance of different frame types, channel conditions, buffer state and the relative start time of the video streams of the users. At the beginning of a time slot the scheduler computes the priorities of all users and schedules the one with the highest priority to transmit. This scheme when compared to non-content aware scheduling ensures that the higher priority frames have a lower frame loss rate. However, it is not clear how to set the priorities assigned to the different frame types, in order to optimize performance. This is particularly an issue when SVC is considered and a larger set of possible priorities exist.

To elaborate on the significance of priority-based content-aware scheduling, a QoE-based packet marking strategy scheduling model is presented in Figure 2. According to the figure, the marking algorithm at the Packet Data Network Gateway (P-GW) provides packet prioritization for video streams having different number of quality enhancement layers. The algorithm at the P-GW exploits the utility functions (based on MOS vs. Bit-rate) of the video streams and mark layers according to their bit-rates and contribution towards the overall perceived video quality. The main goal of the marking is to achieve the maximum video quality under the constraint of the available network resources. Thus, the packets of video layers contributing less towards MOS at the expense of higher bit-rates are marked to be served with lower priority. The higher the priority class, the lower the importance of the marked packets, which is exploited by the scheduler at the eNodeB by dropping such packets when the system becomes highly congested as given in [32]. According to [10,33], priority-based optimized packet marking reduces congestion at the base station and provides timely video rate adaptation at the RAN. However, the approach is limited only to scalable video traffic without considering video traffic types which do not have scalable properties.

Furthermore, to demonstrate the significance of Fuzzy Logic in resource allocation and scheduling, the authors in [34] proposed a novel fuzzy scheduler for cell-edge users in LTE-advanced networks using Voronoi algorithm. In this study, the authors focused on proposing an energy efficient and QoS-aware downlink scheduler for real-time services. Moreover, fuzzy rules were used to optimize the resource allocation for the downlink scheduling of the cell-edge users. The results showed that the proposed scheduler is energy efficient, QoS-aware, and beneficial to the cell-edge users.

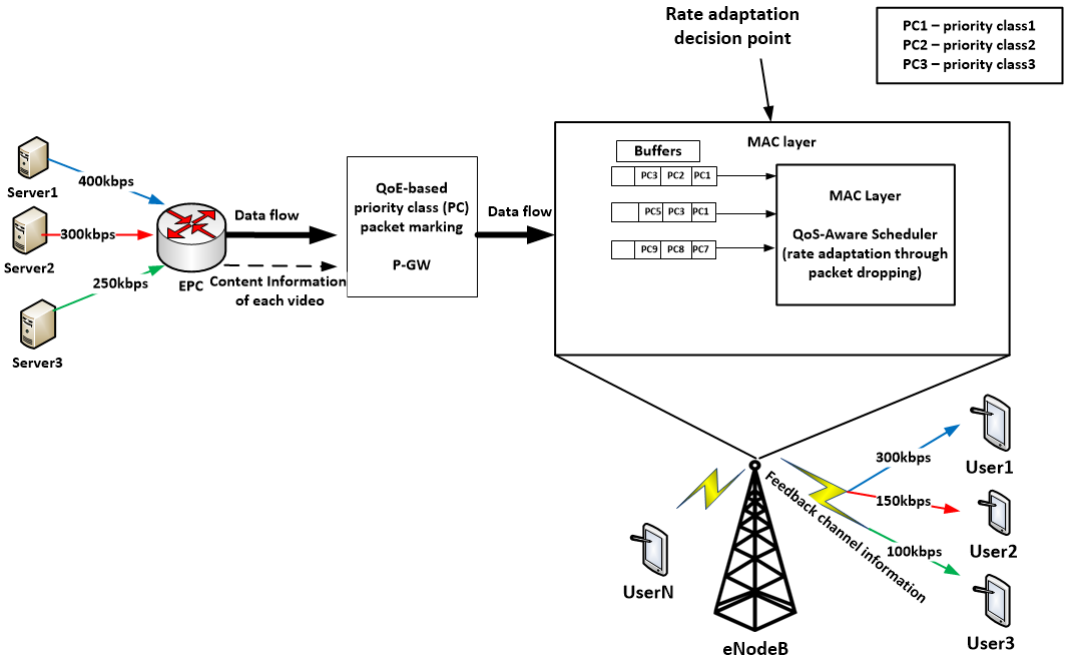

Figure 2. QoE-aware packet priority marking based scheduling.

Moreover, the authors in [35] proposed a joint downlink radio resource allocation and scheduling algorithm for LTE networks using fuzzy-based adaptive priority and effective bandwidth estimation. The resource allocation is based on estimating the utilized bandwidths of traffic flows, whereas the downlink scheduling algorithm is designed to compute the adaptive priorities for the different users by using fuzzy logic. This study focused on ensuring that the QoS parameters are compliant with the requirements of the LTE network. Similarly, the authors in [36] opted for a fuzzy logic approach and proposed a joint scheduling and link adaptation scheme. Furthermore, the proposed scheduler was a priority-based scheme, which optimally allocates radio resources to multiple users based on their QoS requirements for a given application. In addition, the authors went a step further and included power optimisation feature, which would adapt to user's power supply constraints. The results were obtained through numerical evaluations over FIS. The authors concluded that their proposed fuzzy-based scheduler performed similar to the benchmark analytical approach, which utilised Lagrange multipliers scheme, however, with less computational complexity.

In [37], the authors proposed an intelligent fuzzy logic-based channel-aware resource allocation and scheduling scheme over LTE-A networks in the uplink direction. The proposed system was designed to optimally accommodate multi-traffic classes (i.e. Real-time and Non-Real-time). Their channel-aware framework employed Kalman filter controller for channel estimation as well as to meet QoS requirements of the end-users. The performance analysis was carried out in terms of QoS indices (e.g. bandwidth, throughput, fairness, jitter and delay), which indicated that the proposed fuzzy-based scheduler delivered reliable scheduling for real-time services without comprising the non-real-time traffic.

Nevertheless, as mentioned in Section 1, the literature lacks proposals on content-aware priority-based scheduling algorithm utilizing FIS. FIS is beneficial for research and analysis because it provides a trade-off between significance and precision and it relies on concepts of human reasoning that is considered to be most reliable. In addition, the aforementioned studies focus more on content-blind schedulers and lack addressing content-aware scheduling strategies using Fuzzy logic. It is important to note that content-blind schedulers do not produce accurate results as video contents contain different spatio-temporal features, which are their unique signatures, which this paper aims

to address by proposing an intelligent fuzzy logic-based content-aware and channel-aware downlink scheduling algorithm for scalable videos over LTE Networks.

4. Methodology

The methodology followed in this paper can be outlined as proposing a scheduling algorithm based on fuzzy logic, which prioritizes the transmission to multiple users by considering video content, and channel conditions. We start with proposing wireless system model, followed by proposing a fuzzy logic-based content and channel aware downlink scheduler. Next, a novel key performance metric (i.e. significance throughput) is proposed for measuring the performance of the content-aware scheduling algorithm. Lastly, the analytical model is developed in order to analyse the proposed scheduling algorithm.

4.1. The Proposed Wireless System Model

To design the system model, we consider a single 120° sector of a tri-sectored hexagonal cellular downlink orthogonal frequency division multiple access (OFDMA) network, where each cell is served by a base station (BS) with three collocated directional antennas, each serving its respective sector. The tagged sector is serving N active wireless users, uniformly distributed across its area. The system bandwidth for each allocation duration is divided into M physical resource blocks (PRBs), where a PRB is a multi-dimensional resource unit spanning a fixed number of OFDMA subcarriers with bandwidth B and symbol durations.

All the sectors of the same BS can use the same PRBs simultaneously without interference. Adjacent sectors from neighboring cells form a cluster. interference from neighboring clusters is substantially mitigated by the sectorized architecture and the propagation pathloss and fading. A simple inter-cell coordination algorithm is assumed that avoids interference between sectors of the same cluster for all users, by ensuring that neighboring sectors of the same cluster never use the same PRBs simultaneously. By ensuring zero ICI for all users, the number of PRBs required by each user is minimized, thus making available more PRBs for other users in the sector and in the cluster.

The number $m \leq M$ of PRBs allocated to a sector is a function of the mean expected traffic load (in bits/s) requirement of the sector relative to the other sectors in the same cluster. So if L_{cluster} and L_{sector} denote the total expected load of all sectors in the cluster and the load of the tagged sector respectively, then $m \approx L_{\text{sector}}/L_{\text{cluster}}$. The base station can allocate m PRBs to a set of N users at each allocation instance. At each allocation instance multiple PRBs can be assigned to a single user, each PRB however can be assigned to at most one user.

We assume that the channel conditions vary across different PRBs and for different users. The channel conditions vary with time, frequency (e.g. frequency selective multipath fading) and user location. Therefore, each PRB has a corresponding user-dependent and time-varying channel quality that is represented by the maximum possible transmission rate for that user over that PRB. Let $r_i(t,\phi)$ denote the maximum possible transmission rate (bits/s) for user i over PRB ϕ at time t. Then,

$$r_i(t,\phi) = G_{\text{mux}} B \log_2(1 + \epsilon_i(\phi)\gamma_i(t,\phi)), \tag{1}$$

where $\gamma_i(t,\phi)$ is the estimated received signal-to-noise ratio (SNR) after diversity combining (including MIMO antenna diversity and shadow fading), G_{mux} is a MIMO spatial multiplexing gain and $\epsilon_i(\phi)$ is the estimation error margin for $\gamma_i(t,\phi)$. We assume that the channel quality feedback from the user to the BS, comprising of $\gamma_i(t,\phi)$ and $\epsilon_i(\phi)$, are provided to scheduler within the channel's coherence time. Taking bounds of the channel estimation error into account, minimizes the risk of errors during transmission.

4.2. The Proposed Fuzzy Logic-based Content and Channel Aware Downlink Scheduler

Utilising FIS, we propose a downlink scheduling algorithm and a user utility function. In this subsection, we elaborate further on them, respectively.

4.2.1. The downlink scheduling algorithm

The proposed scheduler considers an initial buffer delay or maximum delay constraint T_{D}. Each user must receive one or more GOP, depending on its GOP rate, in this time duration. When a user needs to receive a number g GOPs in this time period, then there are g layers in total with the same index l. The proposed scheduler treats these as a single layer with index l and ensures that they are sent before any of the g layers in total with the same index $l+1$. At the receiver, the layers are re-ordered and reconstituted into frames according to their playback order.

The proposed scheduler, at any time instant, allocates PRBs to users iteratively. Let $\Phi_{\mathrm{ARB}}(t,k)$ and $\Phi_{\mathrm{URB}}(t,k)$ denote the set of allocated and unallocated PRBs, respectively by iteration k of time slot t, Let $\Phi_i(t,k)$ denote the set of PRBs allocated to user i by iteration k of time slot t, and $r_i(t,\varphi)$ is the attainable bit rate of the user on PRB $\varphi \in \Phi_{\mathrm{URB}}(t,k)$. Therefore, each PRB has a corresponding user-dependent and time-varying channel quality that is represented by the maximum possible transmission rate for that user over that PRB. Let $r_i(t,\phi)$ denote the maximum possible transmission rate (bits/s) for user i over PRB ϕ at time t. The expression for $r_i(t,\phi)$ is given in (1).

For each user i, an antecedent layer is sent before any of its descendants. Let $v_i(t,k)$ denote the significance of the layer with the highest significance to be sent to user i by iteration k of time slot t. The user-priority of user i by iteration k of time slot t on PRB $\varphi \in \Phi_{\mathrm{URB}}(t,k)$ is

$$u_i\left(t,k,\varphi\right) = F_{\mathrm{fuzzy}}\left(v_i(t,k), r_i(t,\phi)\right), \tag{2}$$

where the function $F_{\mathrm{fuzzy}}\left(v_i(t,k), r_i(t,\phi)\right)$ is determined by zero-order Sugeno fuzzy inference.

The iterative algorithm operates as follows at any iteration k at time t:

1. For $|\Phi_{URB}\left(t,k\right)| > 0$, find a PRB-user pair which has the highest user utility among all available PRBs and users.
2. $\{i^*, \varphi^*\} = \arg\ \max_{\forall i, \varphi \in \Phi_{\mathrm{URB}}(t,k)} u_i\left(t,k,\varphi\right)$
3. Allocate PRB φ^* to user i^*:
$$\Phi_{\mathrm{PRB},i*}(t,k+1) = \Phi_{\mathrm{PRB},i*}\left(t,k\right) + \{\varphi^*\}$$
4. Delete the PRB from the set of available PRBs:
$$\Phi_{\mathrm{URB}}(t,k+1) = \Phi_{\mathrm{URB}}\left(t,k\right) - \{\varphi^*\}$$
5. Repeat above until all PRBs are allocated, i.e. until $|\Phi_{URB}\left(t,k\right)| = 0$
6. Repeat above steps for new time-slot $t = t + T_{\mathrm{PRB}}$, where T_{PRB} is the time duration of a single PRB.

4.2.2. User utility function based on fuzzy logic

The function defining the user utility $u_i\left(t,k,\varphi\right) = F_{\mathrm{fuzzy}}\left(v_i(t,k), r_i(t,\phi)\right)$ is derived by applying the following fuzzy rule base.

1. *Rule1*: If $significance\ v_i(t,k)$ is *high* then *user_utility* is *high*
2. *Rule2*: If $significance\ v_i(t,k)$ is *low* then *user_utility* is *low*
3. *Rule3*: If *rate* $r_i(t,\phi)$ is *high* then *user_utility* is *high*
4. *Rule4*: If *rate* $r_i(t,\phi)$ is *low* then *user_utility* is *low*

The fuzzy inference is applied to every user i/PRB φ pair for each iteration k of time slot t, where PRB $\varphi \in \Phi_{\mathrm{URB}}(t,k)$.

Let V_{high} and V_{low} denote fuzzy significance sets over the universe of discourse of *significance*, representing *high* and *low significance* set, respectively. Let $U_{\mathrm{V}}(V)$ denote a fuzzy singleton consequent over the universe of discourse of *significance*, where $U_{\mathrm{V}}(V_{\mathrm{high}})$ and $U_{\mathrm{V}}(V_{\mathrm{low}})$ represent the *high* and *low* consequents of *Rule1* and *Rule2*, respectively. Let $\mu_{\mathrm{V}}(v_i(t,k))$ denote the degree of membership or membership function of *significance* with a crisp value $v_i(t,k)$ in the fuzzy set $V \in \{V_{\mathrm{high}}, V_{\mathrm{low}}\}$. Let R_{high} and R_{low} denote fuzzy rate sets over the universe of discourse of *significance*, representing *high* and *low rate* set, respectively. Let $U_{\mathrm{R}}(R)$ denote a fuzzy singleton consequent over the universe of discourse of *rate*, where $U_{\mathrm{R}}(R_{\mathrm{high}})$ and $U_{\mathrm{R}}(R_{\mathrm{low}})$ represent the *high* and *low* consequents of *Rule3* and *Rule4*, respectively. Let $\mu_{\mathrm{R}}(r_i(t,\phi))$ denote the degree of membership or membership function of *rate* with a crisp value $r_i(t,\phi)$ in the fuzzy set $R \in \{R_{\mathrm{high}}, R_{\mathrm{low}}\}$.

Applying a zero-order Sugeno fuzzy inference results in a crisp value for *user_utility* u expressed as

$$u_i(t,k,\varphi) = F_{\mathrm{fuzzy}}(v_i(t,k), r_i(t,\phi)) \equiv \frac{\sum_V \mu_{\mathrm{V}}(v_i(t,k,\varphi))\, U_{rmV}(V) + \sum_R \mu_{\mathrm{R}}(r_i(t,k,\varphi))\, U_{\mathrm{R}}(R)}{\sum_V \mu_{\mathrm{V}}(v_i(t,k,\varphi)) + \sum_R \mu_{\mathrm{R}}(r_i(t,k,\varphi))} \quad (3)$$

The above expression is simplified by selecting membership functions such that $\sum_V \mu_{\mathrm{V}}(v_i(t,k,\varphi)) = 1$ and $\sum_R \mu_{\mathrm{R}}(r_i(t,k,\varphi)) = 1$ then

$$u_i(t,k,\varphi) \equiv 0.5 \left[\sum_V \mu_{\mathrm{V}}(v_i(t,k,\varphi))\, U_{\mathrm{V}}(V) + \sum_R \mu_{\mathrm{R}}(r_i(t,k,\varphi))\, U_{\mathrm{R}}(R) \right] \quad (4)$$

The expression is further simplified by setting $U_{\mathrm{V}}(V) = U_{\mathrm{R}}(R) = 0$ for $V = V_{\mathrm{low}}$ and $R = R_{\mathrm{low}}$.

$$u_i(t,k,\varphi) \equiv 0.5 \left[\mu_{\mathrm{V}}(v_i(t,k,\varphi))\, U_{\mathrm{V}}(V) + \mu_{\mathrm{R}}(r_i(t,k,\varphi))\, U_{\mathrm{R}}(R) \right], \text{for } V = V_{\mathrm{high}} \text{ and } R = R_{\mathrm{high}} \quad (5)$$

Finally, let $U_{\mathrm{V}}(V) = \alpha$ and $U_{\mathrm{R}}(R) = 1 - \alpha$, where $V = V_{\mathrm{high}}$ and $R = R_{\mathrm{high}}$, then:

$$u_i(t,k,\varphi) \equiv 0.5 \left[\mu_{\mathrm{V}}(v_i(t,k,\varphi))\, \alpha + \mu_{\mathrm{R}}(r_i(t,k,\varphi))\, (1-\alpha) \right], \text{for } V = V_{\mathrm{high}} \text{ and } R = R_{\mathrm{high}}, \quad (6)$$

where α is referred to as the *uility coefficient* and determines the trade-off between content and channel awareness.

Linear membership functions are used. The membership functions $\mu_{\mathrm{V}}(v) = v$ and $\mu_{\mathrm{R}}(r) = r/r_{\max}$ for $V = V_{\mathrm{high}}$ and $R = R_{\mathrm{high}}$, respectively, where $r_{\max}$ is the maximum rate in bits/s that can be supported over a single PRB when using the highest order modulation.

4.3. Key System Parameters and Key Performance Metrics

There are two key system parameters for the joint multi-user content and channel aware scheduling, namely the utility coefficient α and the number of users N. Specifically. we consider a single tagged user for observation and $N-1$ competing users. The tagged user is representative of all users within a limited area of the sector. The utility coefficient α determines to what extent the scheduler prioritizes according to channel quality or content importance.

A novel key performance metric is proposed in this paper for evaluating the performance of content-aware scheduling. This metric is the significance throughput $Z_{\mathrm{sig}}(p)$. Other metrics used are the bit throughput $Z_{\mathrm{bit}}(p)$ in bits/s and average PSNR $Q_{\mathrm{P}}(p)$, respectively. The metrics are computed for a tagged user occupying a limited area of the sector containing p% of the closest users to the BS. More elaborations and mathematical expressions on the aforementioned metrics are provided next in Section 4.4.

4.4. The Proposed Analytical Model

There are m PRBs per time-slot, where a time-slot is the allocation period for the scheduler. We consider the allocation over a period of N_{TS} time slots, where the duration of a time-slot is T_{PRB} seconds. The time $T_{\text{PRB}}N_{\text{TS}}$ denotes the maximum delay constraint T_{D} for all layers belonging to a one or more GOPs of a user to be received. The frame rate R_{frame} is related to delay constraint as $R_{\text{frame}} = gN_{\text{frame}}/T_{\text{D}}$, where N_{frame} is the number of frames per GOP and g is the number of GOPs sent in T_{D}. The GOP rate is g/T_{D} and is determined by how the video has been coded. We consider a period of operation of the scheduling algorithm over the duration T_{D}. Figure 3 shows the time and frequency distribution of PRBs over a single allocation period.

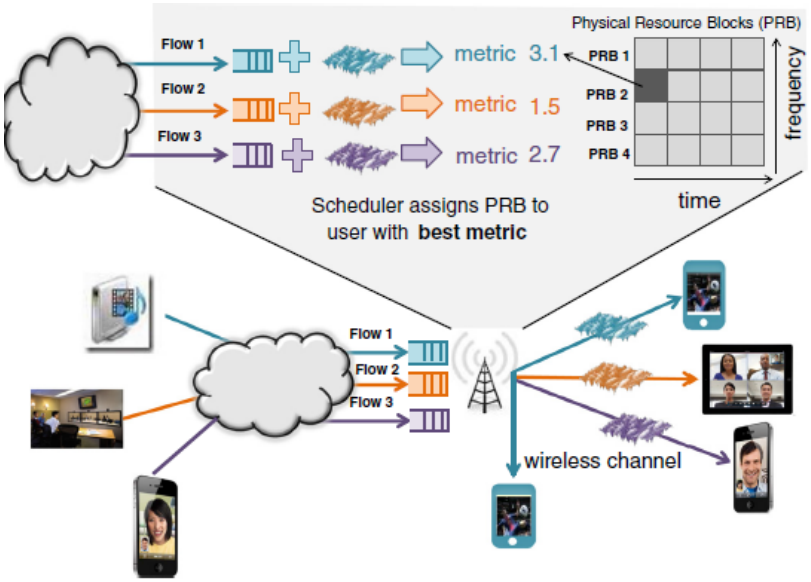

Figure 3. Time/frequency distribution of the PRBs over one allocation period.

We consider a given layer l of a tagged user i competing with the layers of $N-1$ other users. Associated with each layer, after the last iteration of the last slot in this period, is a set of PRBs from the pool of $M_{\text{PRB}} = mN_{\text{TS}}$ PRBs. The set of PRBs associated with each layer represents the minimum number of an arbitrary grouping of PRBs required to send the layer to the user within the T_{D} time period, where feasible, or the total number of PRBs, otherwise. Each PRB from this set falls into two categories. Either the PRB was allocated to the user and used to send bits of the layer or it was not allocated due to competition. Associated with the set of PRBS is an average rate per PRB $R_i(d)$ which is a function of the the user's distance d from the BS, and the long-term shadow fading. It is assumed that short-term frequency selectivity across the set of PRBs is effectively mitigated using diversity, such as frequency domain equalization, interleaving/coding and distributed subcarrier allocation for each PRB. This assumption simplifies the analysis and though it increases the complexity of the physical layer, these suggested diversity schemes are features of advanced OFDMA standards such as LTE. We assume that MIMO is used in antenna diversity mode and not in spatial multiplexing mode, such that $G_{\text{mux}} = 1$.

$$R_i(d) = B\log_2(1 + e_{\text{fade}}\gamma_i(d)), \tag{7}$$

where $\gamma_i(d)$ is the mean SNR and e_{fade} is the shadow fading random variable with pdf $f_{\text{fade}}(e)$. The pdf of $R_i(d)$, $f_{\text{R}}(r,d)$, is obtained as a transformation of $f_{\text{fade}}(e)$.

Users are assumed to be uniformly distributed in the sector, where the radius of the sector is d_{rad} and its area is $A_{\text{sec}} = \pi d_{\text{rad}}^2/3$. The area of the 120°8 sector centered at the BS occupied by p% of the users is $0.01pA_{\text{sec}}$ and has a radius $d_{\text{sec}}(p) = (0.03pA_{\text{sec}}/\pi)^{0.5}$. The parameter $\delta_{\text{user}} = N/A_{\text{sec}}$ is the user density, where N is the total number of users. The distance d of a single user within the sector defined by radius $d_{\text{sec}}(p)$ is a random variable determined by the uniform user distribution with a pdf $f_d(d,p)$, where $d \le d_{\text{sec}}(p) \le d_{\text{rad}}$ and $d_{\text{sec}}(100\%) = d_{\text{rad}}$.

Let $N_{\text{bits}}(l,i)$ denote the number of bits of layer l of user i, which has a probability mass function $P_{\text{LB}}(k,l,i)$. Let $N_{\text{PRB}}(l,i,d)$ denote the number of PRBs required to send layer l of user i, then:

$$N_{\text{PRB}}(l,i,d) = \left\lceil \frac{N_{\text{bits}}(l,i)}{R_i(d)T_{\text{PRB}}} \right\rceil \tag{8}$$

Let $P_{\text{PRB}}(k,l,i,d)$ denote the probability mass function of $N_{\text{PRB}}(l,i,d)$, which is derived from $P_{\text{LB}}(k,l,i)$ and $f_{\text{R}}(r,d)$.

Let $v_{i,l}$ denote the significance of layer l of user i, and let $u_{i,l}(d)$ denote the utility of user i when sending bits of layer l.

$$u_{i,l}(d) = F_{\text{fuzzy}}\left(v_{i,l}, R_i(d)\right), \tag{9}$$

Let $f_{\text{sig}}(l,i,d)$ denote the pdf of $u_{i,l}(d)$, which is derived from $f_{\text{R}}(r,d)$. Let $S_{i,l}(d)$ the number of bits of layer l of user i that are transmitted before the delay constraint. Let i denote th index of the tagged user and l the index of the layer under consideration. Let $\widehat{i}$ denote the index of a competing user and $\widehat{l}$ the index of a layer of this user. Let $S_{\text{PRB}}(l,i,d)$ denote the sum of PRBs required to send the layers of the the competing users that have a higher utility than the tagged user and the layers of he tagged user up to layer l. For $l > 0$

$$S_{\text{PRB}}(l,i,d) = \sum_{\widehat{i}=0,\widehat{i}\neq i}^{N-1} \sum_{\widehat{l}=0}^{N_{\text{L}}(\widehat{i})-1} N_{\text{PRB}}(\widehat{l},\widehat{i},\widehat{d})\text{I}(u_{\widehat{i},\widehat{l}}(\widehat{d}) > u_{i,l}(d)) - \sum_{\lambda=0}^{l-1} N_{\text{PRB}}(\lambda,i,d), \tag{10}$$

where the function $\text{I}(condition)$ equals 0, if $condition$ is $false$ and equals 1, otherwise. For $l = 0$

$$S_{\text{PRB}}(l,i,d) = \sum_{\widehat{i}=0,\widehat{i}\neq i}^{N-1} \sum_{\widehat{l}=0}^{N_{\text{L}}(\widehat{i})-1} N_{\text{PRB}}(\widehat{l},\widehat{i},\widehat{d})\text{I}(u_{\widehat{i},\widehat{l}}(\widehat{d}) > u_{i,l}(d)). \tag{11}$$

The difference $M_{\text{PRB}} - S_{\text{PRB}}(l,i,d)$ determines the number of PRBs available to send layer l of the tagged user. If this difference is zero or negative, then no bits of the layer are sent. If it is non-zero, positive, and less than $N_{\text{PRB}}(l,i,d)$, then some but not all bits of the layer are sent. If it is non-zero, positive, and equal to or more than $N_{\text{PRB}}(l,i,d)$, then all bits of the layer are sent. Therefore,

$$S_{i,l}(d) = \begin{cases} 0 & \text{if } M_{\text{PRB}} - S_{\text{PRB}}(l,i,d) \le 0 \\ R_i(d)\text{E}\left\{ M_{\text{PRB}} - S_{\text{PRB}}(l,i,d) \right\} & \text{if } 0 < M_{\text{PRB}} - S_{\text{PRB}}(l,i,d) \le N_{\text{PRB}}(l,i,d) \\ R_i(d)\text{E}\left\{ N_{\text{PRB}}(l,i,d) \right\} & \text{if } M_{\text{PRB}} - S_{\text{PRB}}(l,i,d) > N_{\text{PRB}}(l,i,d) \end{cases} \tag{12}$$

The significance throughput $S_{\text{sig}}(d)$ of a user at distance d from its BS

$$S_{\text{sig}}(d) = \frac{1}{N_{\text{L}}(i)} \sum_{l=0}^{N_{\text{L}}(i)-1} \frac{S_{i,l}(d)}{R_i(d)\text{E}\{N_{\text{PRB}}(l,i,d)\}}, \tag{13}$$

The bit throughput $S_{\text{bit}}(d)$ of a user at distance d from its BS

$$S_{\text{bit}}(d) = \sum_{l=0}^{N_{\text{L}}(i)-1} \frac{1}{T_{\text{D}}} S_{i,l}(d), \tag{14}$$

The average PSNR $Q(d)$ of a user at distance d from its BS

$$Q(d) = \sum_{l=0}^{N_{\text{L}}(i)-1} (q_l - q_{l-1}) \frac{S_{i,l}(d)}{R_i(d)\text{E}\{N_{\text{PRB}}(l,i,d)\}}, \tag{15}$$

where q_l is the average PSNR if all $l+1$ layers have been received without error and equals zero for $l < 0$.

The significance throughput $Z_{\text{sig}}(p)$ of a user among the p% of users closest to the BS

$$Z_{\text{sig}}(p) = \int_0^{d_{\text{sec}}(p)} f_d(d,p) S_{\text{sig}}(d) \text{d}d, \tag{16}$$

The bit throughput $Z_{\text{bit}}(p)$ of a user among the p% of users closest to the BS

$$Z_{\text{bit}}(p) = \int_0^{d_{\text{sec}}(p)} f_d(d,p) S_{\text{bit}}(d) \text{d}d, \tag{17}$$

The average PSNR $Q_{\text{P}}(p)$ of a user among the p% of users closest to the BS

$$Q_{\text{P}}(p) = \int_0^{d_{\text{sec}}(p)} f_d(d,p) Q(d) \text{d}d. \tag{18}$$

5. Numerical Results and Analysis

All users are assumed to be streaming video sequences with identical traffic and quality statistics. Specifically, statistics of the first hour of the *Tokyo Olympics* video (133 128 frames at 30 frames/sec) [38] are used. Its traffic statistics, quality statistics, and trace are publicly available at [39]. The video sequence is in the Common Intermediate Format (CIF, 352 x 288 pixels). We consider the temporal layers embedded in the video stream encoded with H.264 SVC, with a GOP structure $\{\text{K}_0\text{B}_2\text{B}_1\text{B}_2\text{K}_0..\}$, where K_0 is an I or P key picture. Thus, $N_{\text{L}} = 3$ and $l \in \{0,1,2\}$. The probability distribution $P_{\text{LB}}(k,l,i)$ of layer l and the set of average PSNR values q_l are obtained from [39]. In [39], two values are given for the average PSNR of the key picture, namely, 27.31 dB and 26.94 dB for the I frame and P frame, respectively. In this study, we use the lower value for both types of key pictures and set $q_0 = 26.94$ dB, while q_1 and q_2 equal 28.43 dB and 29.32 dB, respectively.

For the wireless system, parameter values are taken mostly from [40,41]. The channel is assumed to be flat in time and frequency due to the OFDM modulation and the effective exploitation of diversity in the time and frequency domains. Independent lognormal shadow fading with pdf $f_{\text{fade}}(e)$ and a standard deviation of 8 dB has been assumed. The values for the BS antenna gain, UE antenna gain, UE noise figure and total sector TX power are 14 dBi, 0 dBi, 7 dB and 46 dBm, respectively. The time-slot duration T_{PRB} is 0.5 ms. The coverage of the sector has a radius of 250m. The maximum number of PRBs per sector m is 34 per slot, with each PRB having a bandwidth of 180 kHz. Furthermore, we assume a maximum spectral efficiency of 6 bits/s/Hz in each PRB, for 64QAM modulation without MIMO spatial multiplexing. Therefore, the maximum bit rate per PRB r_{max} is 1080 kbits/s. A 2x2 MIMO antenna diversity gain of 6 dB is assumed. The distance-dependent path gain is given by $-128.1 - 37.6 \log_{10}(d)$.

Of all the statistical distributions used in the analytical model, the distributions $P_{\text{LB}}(k,l,i)$ and $f_{\text{fade}}(e)$ are all given, the former is obtained empirically [39], while the latter is assumed to be

Table 1. Table of symbols.

$Q(d)$, $S_{\text{sig}}(d)$, $S_{\text{bit}}(d)$	Average PSNR, Significance throughput and Bit throughput, respectively, for a user at distance d from the BS
$Q_{\text{P}}(p)$, $Z_{\text{sig}}(p)$ and $Z_{\text{bit}}(p)$	Average PSNR, Significance throughput and Bit throughput, respectively, of a user among the p% of users closest to the BS
q_l	Average PSNR if all $l+1$ layers have been received without error
$S_{i,l}(d)$	Number of bits of layer l of user i that are transmitted before the delay constraint.
$v_{i,l}$	Significance of layer l of user i
$R_i(d)$, $f_{\text{R}}(r,d)$	Average rate per PRB $R_i(d)$ and its pdf as a function of the the user's distance d from the BS
$u_{i,l}(d)$	Utility of user i when sending bits of layer l.
$f_{\text{sig}}(l,i,d)$	The pdf of $u_{i,l}(d)$
$N_{\text{PRB}}(l,i)(d)$	Number of PRBs required to send layer l of user i
$N_{\text{bits}}(l,i)$	Number of bits of layer l of user i,
d_{rad}	Radius of cell sector
A_{sec}	Area of cell sector
$d_{\text{sec}}(p)$	Radius of a sector centered at the BS occupied by p% of the users
$f_d(d,p)$	The pdf of the distance d of a single user within the sector defined by radius $d_{\text{sec}}(p)$.
T_{PRB}	The time duration of a single PRB or time-slot
R_{frame}	Frame rate
N_7	Number of frames per GOP
N_{TS}, M_{PRB},T_{D}	Maximum number of time-slots, maximum number of PRBs and maximum delay constraint or duration, respectively, to send all layers of a GOP
m	Number of PRBS per time-slot
i,l	Indexes of the tagged user and tagged layer, respectively.
$\widehat{i},\widehat{l}$	Indexes of a competing user and layer, respectively.
$\gamma_i(d)$	Mean SNR of a user at distance d from the BS.
α	Utility coefficient
$N_{\text{PRB}}(l,i,d)$,$P_{\text{PRB}}(k,l,i,d)$	Number of PRBs required to send layer l of user i, and its probability mass function, respectively
$N_{\text{bits}}(l,i)$, $P_{\text{LB}}(k,l,i)$	Number of bits of layer l of user i and its probability mass function, respectively
N	Number of users
e_{fade}, $f_{\text{fade}}(e)$	Shadow fading random variable and its pdf, respectively.

lognormal. The pdf of $f_d(d,p)$ is derived from a transformation, given that d^2 is uniformly distributed over the range $(0, d_{\text{sec}}(p)]$. The other distributions mentioned above are derived from one or more of these three distributions, and are obtained numerically from Monte-Carlo simulations.

Figure 4 shows a plot of PSNR versus number of users for different classes of users and for $\alpha = 0$, where users are classified according to the region they occupy in the sector. This scenario corresponds to channel-aware only scheduling. The results show that the closest 20% of users acheive the maximum PSNR performance over the entire observed range. The PSNR deteroriates at a rapid rate the further the range of users considered.

Figure 5 shows the corresponding plot using the significance throughput metric for different classes of users and for $\alpha = 0$. The results show that the closest 20% of users acheive the maximum significnace throughput of unity, hence the maximum PSNR performance, over the entire observed range. A frame rate of 30fps is considered, which maps to a maximum delay constraint of 0.1333 secs to deliver all the layers comprising the the 4 frames of the GOP. The significance throughput follows the same trend as the PSNR and deteroriates at a rapid rate the further the range of users considered.

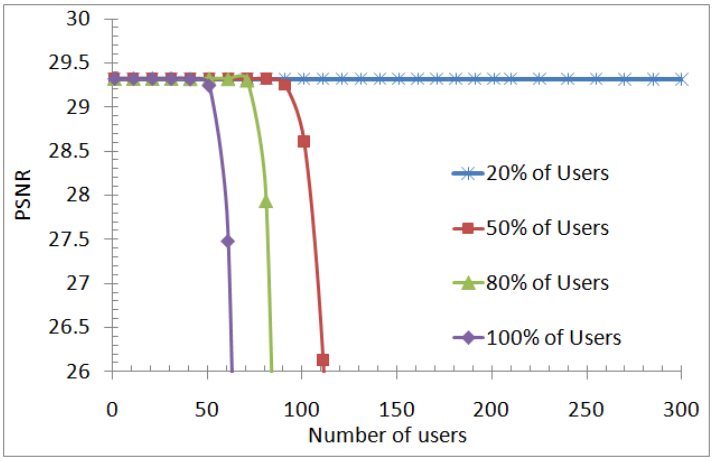

Figure 4. PSNR versus number of users for different classes of users and $\alpha = 0$.

Figure 6 and Figure 7 shows corresponding plots using the significance throughput for $\alpha = 0.25$ and

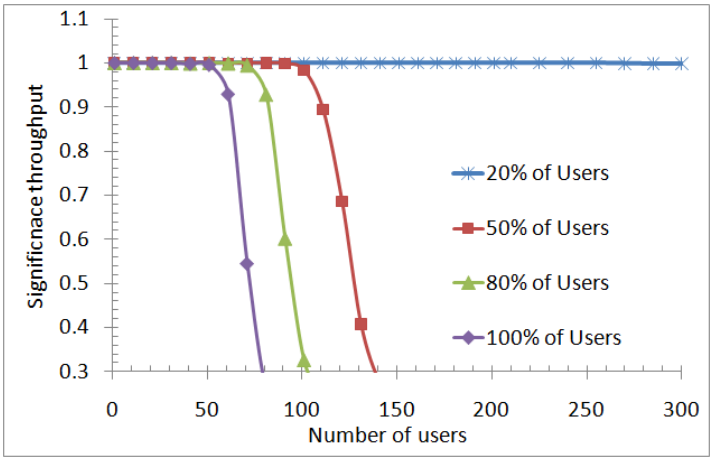

Figure 5. Significance throughput versus number of users for different classes of users and $\alpha = 0$.

$\alpha = 0.5$, respectively. These plots show the impact of introducing content-awareness, while reducing proportionately the extent of channel awareness. The results show that the performance of the closest 20% of users declines at a rapid rate as α is increased from zero to 0.5, while the performance of the users measured over larger distance ranges improves at a slower rate. For $\alpha = 0.5$, the perfoamnce across all distance ranges have converged signficantly.

The objective of introducing content awareness is to improve the fairness of the proposed scheduling algorithm, in general, and particularly to enhance the performance of distant users at a minimum penalty to users close to the BS. For illustration, we consider minimum significance throughput target for the closest 20% of users to be $\frac{2}{3}$, which corresponds to receiving the most important two out of the three layers. For the closest 100% of users, that is all users, we consider minimum significance to be $\frac{1}{3}$, which corresponds to receiving the most important one out of the three layers. With these constarints, the maximum number of users that can be supported increases from 78% to 90% (a 15% enhancement), as α increases from 0 to 0.25. It declines to 72% as α increases from 0.25 to 0.5.

Figure 8, Figure 9 and Figure 10 show plots of significance throughput versus number of users for different maximum delay constraints, corresponding to slightly reduce frame rates. Since the playback rate is constant at 30fps, reducing the frame-rate at scheduler implies that the video sequence will experience short pauses. the shorter the pauses the less perceptible to the user. Figure 8

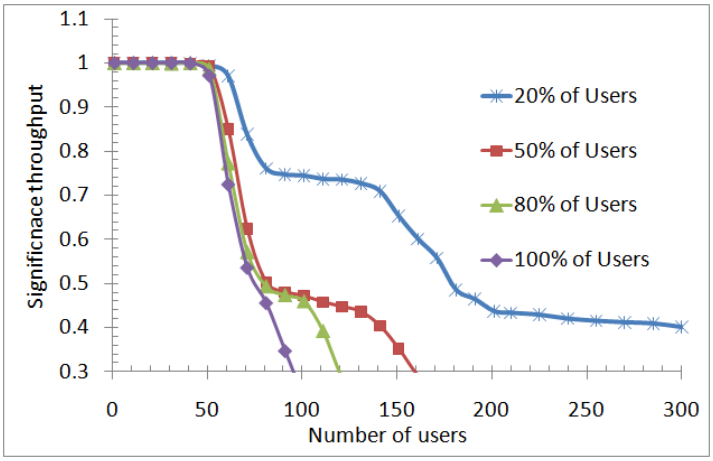

Figure 6. Significance throughput versus number of users for different classes of users and $\alpha = 0.25$.

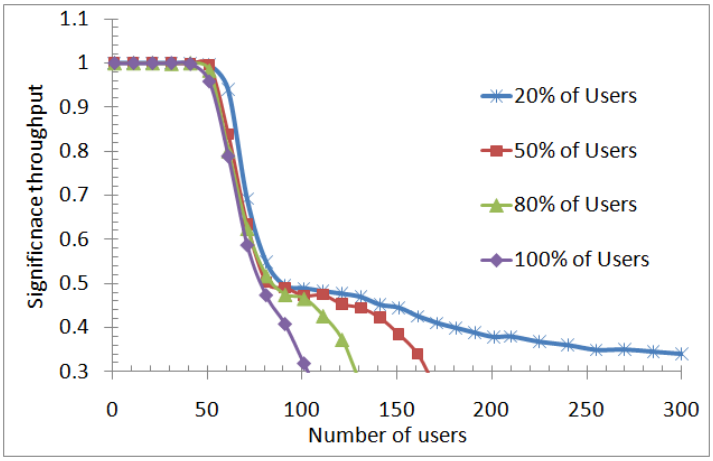

Figure 7. Significance throughput versus number of users for different classes of users and $\alpha = 0.5$.

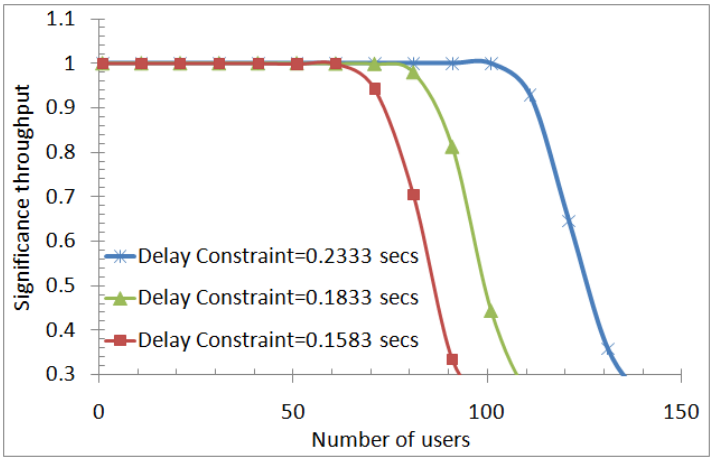

Figure 8. Significance throughput versus number of users for different maximum delay constraints, 100%of users and $\alpha = 0$.

and Figure 10 show the case for 100% of the users with α equal to zero and 0.25, respectively, while Figure 9 shows the case for 20% of the users and $\alpha = 0.25$.

The results show that increasing the delay constraint increases the significance throughput, and hence PSNR, for all cases. The results show that significant performance improvement is possible for small increases in delay, or equivalently small reductions in frame rate. Consider an increase in the delay constraint from 0.1333 secs to 0.1583 secs, corresponding to a reduction in frame rate at the scheduler from 30fps to 25.27fps. Given the constraints on the minimum significance throughput for the closest 20% and 100% users mentioned above, the maximum number of users that can be supported increases from 91% to 109% (a 19.8% enhancement), as α increases from 0 to 0.25.

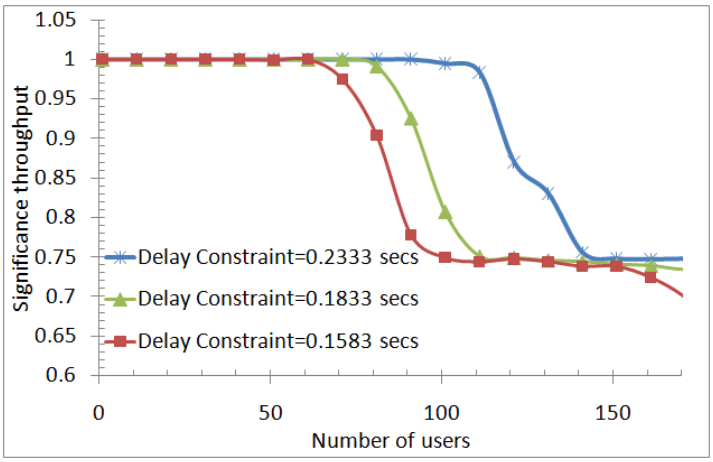

Figure 9. Significance throughput versus number of users for different maximum delay constraints, 20% of users and $\alpha = 0.25$.

Table 2. Simulation results for content-aware and content-unaware scheduling strategies.

Parameters	Fuzzy-based Scheduler					PF-Scheduler					M-LWDF-Scheduler				
Channel Aware	x					x					x				
Delay Aware											x				
Content Aware	x														
Results															
Metrics	Network Load (Users)					Network Load (Users)					Network Load (Users)				
	8	12	16	20	24	8	12	16	20	24	8	12	16	20	24
PSNR (dB)	35.8	34.9	32	30.7	29	31.8	25.3	21.5	20.8	20.8	37.1	30.8	21.5	20.8	20.8

Table 2 shows the simulation results for content-aware and content-unaware scheduling strategies. A comparison between the proposed fuzzy-based scheduler and the standard schedulers has been carried out, as these schedulers provide fairly good performance in terms of PSNR and number of users. The simulation framework and the channel model parameters are the same as in our previous study [2]. Furthermore, we also select the same SVC videos as in our previous study [2]. The SVC layers of different video contents are marked with a priority index according to the QoE based marking algorithm in [11]. As a benchmark strategy, we utilize the Proportional Fair (PF) and M-LWDF schedulers. The simulation results of the proposed and benchmark strategies are reported in Table 2. According to the table, the proposed fuzzy based scheduler achieves a cumulative video quality of 35.8 dB when the total number of video streaming users is 8. On the other hand, PF and M-LWdF schedulers achieves a video quality of 31.8 dB and 37.1 dB respectively. The increase in load (in terms of the total number of streaming users) decreases the cumulative video quality of all the strategies. However, the degradation in video quality of the proposed fuzzy-based scheduler is lower as compared to the benchmark strategies. This is mainly because the fuzzy-based scheduler

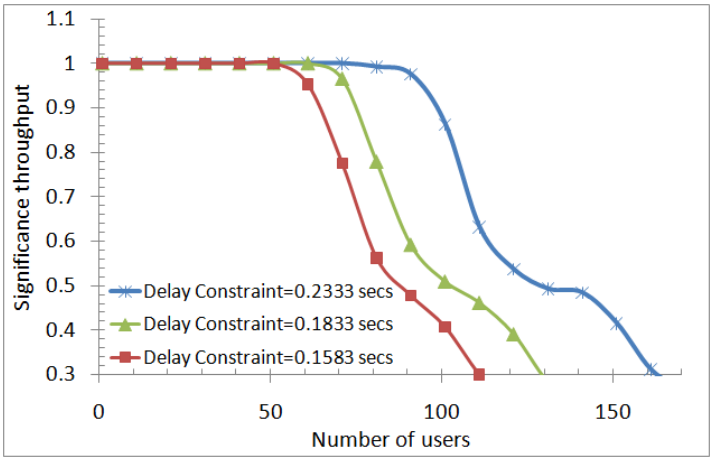

Figure 10. Significance throughput versus number of users for different maximum delay constraints, 100% of users and $\alpha = 0.25$.

prioritizes the most important SVC layers. Therefore, layers contributing highest to the QoE are scheduled before their deadline. The increase in load prioritizes the most important SVC layer of each user. On the other hand, PF and M-LWDF strategies assign radio resources to the SVC layers irrespective of their quality contributions, which results in a steep fall in the quality, as shown in Table 2, when the number of users is increased from 8 to 24.

6. Conclusions

A novel intelligent fuzzy logic-based content and channel aware downlink scheduler for scalable video streaming has been proposed. Using novel content-aware and standard performance metrics, the performance of the proposed scheduling algorithm through the analytical model is evaluated. The fuzzy logic controller allows a single utility parameter to be defined and use the trade-off between content and channel-awareness in order to enhance the overall user experience throughout the coverage area. The results show that the number of supported users can be enhanced by as much as 15%, for playback without pauses and as much as 19% if short imperceptible pauses are acceptable. Significantly, the results demonstrate that channel-aware only and content-aware only schemes are inadequate for supporting video services in a cellular environment. The former delivers disproportionately good quality to users close to the BS, while users at the sector edge are unable to meet a minimum quality. The latter significantly penalises users with good channels, while the performance of edge users, though improved, remains minimal. The proposed algorithm allows for a fairer allocation of resources among users across the entire sector coverage, allowing for the enhancement of video quality at edges of the cell while minimizing the degradation to users closer to the BS. Future work will consider heterogenous video traffic and the sensitivity to different fuzzy rule bases and membership functions for the fuzzy-controller. In addition, a performance analysis between our proposed scheduling algorithm and other content and channel aware scheduling algorithms will be considered.

Acknowledgments: Dr. Peter Omiyi would like to acknowledge the Holon Institute for its support. Also, Dr. Moustafa Nasralla would like to acknowledge the management of Prince Sultan University (PSU) and the Renewable Energy Lab for the valued support and research environmental provision which have led to completing this work. Moreover, Dr. Ikram Ur Rehman would like to acknowledge the West London University for its support. Finally, Prof. Maria Martini would like to acknowledge Kingston University for its support.

Conflicts of Interest: The authors declare no conflict of interest.

Bibliography

1. Nasralla, M.M. A Hybrid Downlink Scheduling Approach for Multi-Traffic Classes in LTE Wireless Systems. *IEEE Access* **2020**, *8*, 82173–82186.
2. Nasralla, M.M.; Khan, N.; Martini, M.G. Content-aware downlink scheduling for LTE wireless systems: A survey and performance comparison of key approaches. *Computer Communications* **2018**, *130*, 78 – 100.
3. Nasralla, M.M.; Razaak, M.; Rehman, I.U.; Martini, M.G. Content-aware packet scheduling strategy for medical ultrasound videos over LTE wireless networks. *Computer Networks* **2018**, *140*, 126–137.
4. Rehman, I.U.; Nasralla, M.M.; Philip, N.Y. Multilayer perceptron neural network-based QoS-aware, content-aware and device-aware QoE prediction model: a proposed prediction model for medical ultrasound streaming over small cell networks. *Electronics* **2019**, *8*, 194.
5. Xu, Z.; Cao, Y.; Wang, W.; Jiang, T.; Zhang, Q. Incentive Mechanism for Cooperative Scalable Video Coding (SVC) Multicast Based on Contract Theory. *IEEE Transactions on Multimedia* **2019**.
6. Ghermezcheshmeh, M.; Shah-Mansouri, V.; Ghanbari, M. Analysis and performance evaluation of scalable video coding over heterogeneous cellular networks. *Computer Networks* **2019**, *148*, 151–163.
7. van der Schaar, M.; Andreopoulos, Y.; Hu, Z. Optimized Scalable Video Streaming over IEEE 802.11a/e HCCA Wireless Networks under Delay Constraints. *IEEE Trans. on Mobile Computing* **2006**, *5*, 755–768.
8. Pahalawatta, P.V.; Berry, R.; Pappas, T.N.; Katsaggelos, A.K. Content-aware resource allocation and packet scheduling for video transmission over wireless networks. *IEEE Journal on Sel. Areas on Commun.* **2007**, *25*, 749–759.
9. Martini, M.G.; Tralli, V. Video quality based adaptive wireless video streaming to multiple users. Proc. IEEE International Symposium on Broadband Multimedia Systems and Broadcasting; , 2008; pp. 1–4.
10. Fu, B.; Staehle, D.; Kunzmann, G.; Steinbach, E.; Kellerer, W. QoE-aware Priority Marking and Traffic Management for H.264/SVC-based Mobile Video Delivery. 8th ACM workshop on Performance monitoring and measurement of heterogeneous wireless and wired networks; , 2013; pp. 173–180.
11. Fu, B.; staehle, D.; Kunzmann, G.; Steinbach, E.; Kellerer, W. QoE-based SVC layer dropping in LTE networks using content-aware layer priorities. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* **2015**, *12*, 23.
12. Zhang, Y.; Liu, G. Fine granularity resource allocation algorithm for video transmission in orthogonal frequency division multiple access system. *IEEE IET (Institution of Engineering and Technology) Communications* **2013**, *7*, 1383–1393.
13. Li, F.; Ren, P.; Du, Q. Joint Packet Scheduling and Subcarrier Assignment for Video Communications Over Downlink OFDMA Systems. *IEEE Transactions on Vehicular Technology* **July. 2012**, *61*, 2753–2767.
14. Li, F.; Zhang, D.; Wang, M. Multiuser multimedia communication over orthogonal frequency-division multiple access downlink systems. *Concurrency and Computation: Practice and Experience* **2013**, *25*, 1081–1090.
15. Li, P.; Chang, Y.; Feng, N.; Yang, F. A Cross-Layer Algorithm of Packet Scheduling and Resource Allocation for Multi-User Wireless Video Transmission. *IEEE Transactions on Consumer Electronics* **Sept. 2011**, *57*, 1128–1134.
16. Ji, X.; Huang, J.; Chiang, M.; Lafruit, G.; Catthoor, F. Scheduling and Resource Allocation for SVC Streaming over OFDM Downlink Systems. *IEEE Trans. on Circuits and Systems* **2009**, *19*, 1549–1555.
17. Cicalò, S.; Tralli, V. Distortion-Fair Cross-Layer Resource Allocation for Scalable Video Transmission in OFDMA Wireless Networks. *IEEE Transactions on Multimedia* **April. 2014**, *16*, 848–863.
18. Blej, M.; Azizi, M. Comparison of Mamdani-type and Sugeno-type fuzzy inference systems for fuzzy real time scheduling. *International Journal of Applied Engineering Research* **2016**, *11*, 11071–11075.
19. Kaur, A.; Kaur, A. Comparison of Mamdani-type and Sugeno-type fuzzy inference systems for air conditioning system. *International journal of soft computing and engineering* **2012**, *2*, 323–325.
20. Sivanandam, S.; Sumathi, S.; Deepa, S.; others. *Introduction to fuzzy logic using MATLAB*; Vol. 1, Springer, 2007.
21. Jang, J.S. ANFIS: adaptive-network-based fuzzy inference system. *IEEE transactions on systems, man, and cybernetics* **1993**, *23*, 665–685.
22. Schwarz, H.; Marpe, D.; Wiegand, T. Overview of the scalable video coding extension of the H.264/AVC standard. *IEEE Transactions on Circuits and Systems for Video Technology* **2007**, *17*, 1103–1120.

23. Khan, N.; Nasralla, M.M.; Martini, M. Network and User Centric Performance Analysis of Scheduling Strategies for Video Streaming over LTE. IEEE International Conference on Communications (ICC) - Workshop on Quality of Experience-based Management for Future Internet Applications and Services (QoE-FI); , 2015.
24. Khan, N.; Martini, M.G. QoE-driven multi-user scheduling and rate adaptation with reduced cross-layer signaling for scalable video streaming over LTE wireless systems. *EURASIP Journal on Wireless Communications and Networking* **2016**, *2016*, 93.
25. Maani, E.; Pahalawatta, P.V.; Berry, R.; Katsaggelos, A.K. Content-aware packet scheduling for multiuser scalable video delivery over wireless networks. SPIE Optical Engineering and Applications; , 2009.
26. Pahalawatta, P.; Berry, R.; Pappas, T.; Katsaggelos, A. Content-aware resource allocation and packet scheduling for video transmission over wireless networks. *IEEE Journal on Selected Areas in Communications* **2007**, *25*, 749–759.
27. Khan, N.; Martini, M.G.; Bharucha, Z. Quality-aware fair downlink scheduling for scalable video transmission over LTE systems. IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC); , 2012; pp. 334–338.
28. Khan, N.; Martini, M.G.; Staehle, D. Opportunistic Proportional Fair Downlink Scheduling for Scalable Video Transmission over LTE Systems. IEEE Vehicular Technology Conference (VTC); , 2013.
29. Mostafa, A.E.; Gadallah, Y. A statistical priority-based scheduling metric for M2M communications in LTE networks. *IEEE Access* **2017**, *5*, 8106–8117.
30. Zhang, W.; Ye, S.; Li, B.; Zhao, H.; Zheng, Q. A priority-based adaptive scheme for multi-view live streaming over HTTP. *Computer Communications* **2016**, *85*, 89–97.
31. Edelman, B.A.; Gay, J.; Lozben, S.; Shetty, P. Real-time priority-based media communication, 2018. US Patent 9,900,361.
32. Khan, N.; Martini, M. Hysteresis based Rate Adaptation for Scalable Video Traffic over an LTE Downlink. IEEE International Conference on Communications (ICC) - Workshop on Smart Communication Protocols and Algorithms; , 2015.
33. Khan, N. Quality-Driven Multi-User Resource Allocation and Scheduling Over LTE for Delay Sensitive Multimedia Applications. In *Ph.D. Dissertation*; Kingston University London, UK, 2014.
34. Radhakrishnan, S.; Neduncheliyan, S.; Thyagharajan, K. A novel fuzzy scheduler for cell-edge users in LTE-advanced networks using Voronoi algorithm. *Cluster Computing* **2017**, pp. 1–11.
35. Abrahão, D.C.; Vieira, F.H.T. Resource allocation algorithm for LTE networks using fuzzy based adaptive priority and effective bandwidth estimation. *Wireless Networks* **2018**, *24*, 423–437.
36. Taki, M.; Heshmati, M.; Omid, Y. Fuzzy-based optimized QoS-constrained resource allocation in a heterogeneous wireless network. *International Journal of Fuzzy Systems* **2016**, *18*, 1131–1140.
37. Mardani, M.R.; Ghanbari, M. Robust resource allocation scheme under channel uncertainties for LTE-A systems. *Wireless Networks* **2019**, *25*, 1313–1325.
38. der Auwera, G.V.; David, P.T.; Reisslein, M.; Karam, L.J. Traffic and Quality Characterization of the H.264/AVC Scalable Video Coding Extension. *Eurasip Journal on Advances in Multimedia* **2008**, *25*.
39. http://trace.eas.asu.edu/. H.264/AVC and SVC Video Trace Library.
40. 3GPP TR 25.814 V7.1.0. Physical Layer Aspects for Evolved Universal Terrestrial Radio Access (UTRA) (Release 7), 2006.
41. 3GPP TS 36.211 V8.2.0. EUTRA Physical Channels and Modulation (Release 8), 2008.