Content-aware downlink scheduling for LTE wireless systems: A survey and performance comparison of key approaches

Moustafa M. Nasralla, Nabeel Khan, Maria G. Martini

PII: S0140-3664(18)30202-0
DOI: https://doi.org/10.1016/j.comcom.2018.08.009
Reference: COMCOM 5765

To appear in: *Computer Communications*

Received date : 7 March 2018; Revised date : 7 July 2018; Accepted date : 18 August 2018

Content-Aware Downlink Scheduling for LTE Wireless Systems: A survey and performance comparison of key approaches

Moustafa M. Nasralla[1], Nabeel Khan[2], and Maria G. Martini,[2]

[1]Department of Communications and Networks Engineering, Prince Sultan University, Saudi Arabia
[2]Department of Science, Engineering and Computing, Kingston University London, UK

***Abstract*—We present in this paper a comprehensive review and comparison of recent downlink scheduling approaches for video streaming traffic over the Orthogonal Frequency Division Multiple Access (OFDMA) based Long-Term Evolution (LTE) wireless technology. Focusing on content-aware downlink scheduling approaches, we provide an extensive literature review, a taxonomy for content-aware and content-unaware downlink schedulers, and tables that summarize the key approaches and common parameters among the schedulers. In addition, we analyze and compare via simulation the performance of some of the most relevant scheduling rules. Our main goal is to compare and analyze different classes of scheduling strategies in terms of network centric performance metrics as well as user centric metrics. Quality of Service (QoS) evaluation involves the evaluation of network performance parameters, e.g., packet loss rate, average system throughput and end-to-end packet delay. On the other hand, Quality of Experience (QoE) reflects the user's experience and satisfaction in terms of Mean Opinion Score (MOS). According to simulation results, proxy based QoE aware scheduling strategies perform best in terms of number of satisfied users and should be used in an LTE downlink to offer high quality video streaming services.**

***Index Terms*—Downlink scheduling approaches, Content-aware, QoS, QoE, Scheduling optimization, LTE, and OFDMA.**

I. Introduction

The evolution of the 4th generation (4G) wireless technologies as well as 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 78% of the global consumer traffic by 2021 [1]. This has increased the challenge on mobile operators towards designing downlink packet scheduling and resource allocation algorithms for multimedia traffic. Moreover, the delivery of contents for a variety of personalized services over wireless networks is a crucial task to handle. In recent years, much effort has been put on proposing several downlink packet scheduling and radio resource allocation approaches, with the goal of maximizing the QoS and the QoE for the end-users.

Quality of Service (QoS) is a network centric performance assessment approach which involves the evaluation of network performance parameters, e.g., packet loss rate, average system throughput, end-to-end packet delay, system efficiency, and fairness [2]. QoS-aware scheduling approaches take into account the aforementioned network performance parameters in order to utilize the resources of the wireless systems efficiently, schedule packets reliably and produce robust network performance parameters which will indicate a good QoS delivery to the end-users.

On the other hand, Quality of experience (QoE) is a user centric performance assessment approach which reflects the user's experience and satisfaction for the service used [3]. QoE evaluation can be performed via subjective tests with the help of a panel of human observers, in order to obtain MOS which reflects the quality perceived by the observers, and is dependent on human visual system (HVS) [4]. Since subjective tests are time demanding and costly, objective metrics have been developed to estimate the user's perceived quality. These are mathematical based metrics for video quality assessment, such as Mean Square Error (MSE) [5], Peak Signal to Noise Ratio (PSNR) [5] and Structural Similarity (SSIM) [6]. A study on subjective and objective video quality assessment metrics is available for instance in [7–10] and a good and thorough review is provided in [11]. Moreover, a comprehensive survey of the evolution of video quality assessment methods, analyzing their characteristics, advantages, and drawbacks are introduced in [11]. In addition, an introduction to QoE-based video applications is presented.

QoE aware scheduling approaches consider users' satisfaction, with the help of the aforementioned objective and subjective based parameters, in order to utilize the resources of the wireless systems efficiently, schedule packets reliably and maximize the users' perceived video quality. Some examples of research works that considered and proposed QoE assessment methods for video services using different scheduling approaches in LTE networks are presented in [12–15]. Moreover, an extensive literature review on issues related to QoE and its recent applications in video transmission, with consideration of the compelling features of QoE (i.e., context and human factors) is presented in [16]. The issues related to QoE include QoE modeling with influencing factors in the end-to-end chain of video transmission, QoE assessment (including subjective test and objective QoE monitoring) and QoE management of video transmission over different types of networks. The authors also highlight the significance of the context and human factors in QoE-aware video transmission to current research.

Content-awareness [17] is a key-feature to provide QoE. Content-aware scheduling approaches [17] take into account the content of the video streams requested by users. The video stream comprises several frames wherein each frame is fragmented into several packets for transmission. The importance of each video packet can be determined/marked and packets can be ordered according to their relative contribution towards the overall perceived video quality (e.g., distortion, PSNR, Video Quality Metric (VQM) and SSIM). Further, a content-aware utility function can be defined and used for the scheduling decision.

A review of content-aware resource allocation schemes for video transmission over wireless networks is provided in [17], although this does not include the most recent scheduling approaches and wireless technologies; a general review on downlink scheduling approaches for LTE wireless systems is provided in [18], although this only mentions content-unaware approaches; a survey on emerging concepts and challenges for QoE management of multimedia services are discussed in [19] where this study focuses mainly on QoE modeling rather than packet scheduling approaches; the literature review in [20] focuses on relevant radio interference and resource management (RIRM) schemes that have been proposed in the last few years toward 5G radio access networks, although the addressed issues in this review only include discussion about interference, spectrum-efficient, and energy-efficient management schemes and not specifically about content-aware based management.

In this paper, a comprehensive literature review of downlink content-aware scheduling and radio resource allocation strategies over wireless networks has been conducted, discussing the state-of-the-art solutions and approaches in this scope. In order to provide a better understanding of the conducted work, we propose a taxonomy classifying the scheduling algorithms into two main classes, namely content-aware and content-unaware scheduling approaches, each containing further classes of approaches. In addition, scheduling approaches, techniques and common parameters among the important schedulers are listed in tables based on the reviewed literature. Results and comparisons among the classified state-of-the-art strategies are also reported. The work presented in this paper is novel as it addresses the topic of downlink packet scheduling comprehensively, which is lacking in the literature. Few studies have contemplated on downlink scheduling approaches for LTE wireless networks. However, these studies do not include the most recent scheduling approaches and are mostly content blind (i.e. content-unaware).

The rest of the paper is organized as follows: An overview of LTE wireless systems and downlink packet scheduling is presented in Section II. Section III compares and contrasts the reviewed approaches of content-aware schedulers, where we provide tables and figures to differentiate between the reviewed categories and to show their different approaches. Section IV discusses the standardization efforts by the Third Generation Partnership Project (3GPP) for QoE-aware video delivery. Section V reports and compares the performance of the different strategies via results analysis. Finally, Section VI concludes the paper.

II. LTE Wireless Systems and Downlink Packet Scheduling Overview

This section presents a brief overview of LTE wireless networks, the main concepts of different downlink packet scheduling algorithms, and the proposed taxonomy.

A. The LTE wireless system

LTE has been introduced by the 3GPP as the next technology after the 3.5G (HSPA+) cellular networks. Evolved NodeB (eNodeB) is the central authority in the Radio Access Network (RAN) where Radio Resource Management (RRM) and packet scheduling process is performed. LTE uses OFDMA in the downlink transmission mode and Single Carrier Frequency Division Multiple Access (SC-FDMA) in the uplink transmission mode. OFDMA extends the multi-carrier technology Orthogonal Frequency Division Multiplexing (OFDM) to provide a better and more flexible multiple access scheme. In other words, OFDMA splits the frequency band into multiple orthogonal sub-carriers. This helps in improving the system capability to support high data-rates, provide multi-user diversity, and reduce the Inter-Symbol-Interference (ISI) [21–23]. The basic resource unit in LTE is called Physical Resource Block (PRB), equal to 180 KHz bandwidth in the frequency domain and 0.5ms duration in the time domain. Furthermore, the allocated PRBs in OFDMA are flexible in position when assigning them to different users wherein each PRB is modulated by different data symbols depending on their channel quality. For example, if 64 Quadrature Amplitude Modulation (QAM) is employed, each sub-carrier may carry 6 bits of data represented by one of the 64 possible symbols of the 64 QAM in ideal channel conditions. In this case, the 12 sub-carriers will carry data length up to $(12 \times 6 = 72)$ bits for a duration of LTE Symbol (71.4 μsec). The LTE standard comprises a scalable range of channel bandwidth sizes, as each bandwidth size is composed of different number of PRBs. The efficient use and proper selection of these bandwidths will improve the system efficiency and avoid wasting radio resources as studied in [24].

Channel Quality Indicator (CQI) feedback represents the users' instantaneous channel quality at each PRB which can be reported to the eNodeB by the users using uplink control messages over the Physical Uplink Shared Channel (PUSCH). The feedback mechanism functions in a way that each User Equipment (UE) sends a single CQI about every PRB in the corresponding channel bandwidth to the eNodeB. The Signal to Interference Noise Ratio (SINR) and CQI values are mapped in accordance with the mapping tables presented in [25]. As a result, the selected Modulation and Coding Scheme (MCS) guarantees a robust communication and good service delivery. Moreover, this decision ensures that the estimated Block Error Rate (BLER) remains under the target BLER of 10% [25], [26].

B. *Downlink scheduling overview*

OFDMA is employed in LTE as access strategy in downlink transmission, hence the existence of dynamic management of radio resource allocation is very important. Scheduling and resource allocation are responsible for controlling and assigning the PRBs among flows at the Medium Access Control (MAC) layer of the eNodeB. Due to the importance of the scheduling and resource allocation process in today's communications, 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 Proportional Fair (EXP-PF) schedulers [27]. With regard to the previously mentioned schedulers, each flow is assigned a priority value depending on specific measurements. Therefore, 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 signals to multiple users over wireless systems (refer to the example scenario in Figure 2) a scheduling strategy should address the trade-off between resource utilization and fairness among users. On the one hand, the interest of network operators is to maximize the exploitation of the resources, e.g., assigning more resources to the user(s) experiencing better channel conditions. On the other hand, this strategy can result in unsatisfied users, since users experiencing worse channel conditions would not be served and would not meet their QoS and QoE requirements. We will show in the subsequent sections how this trade off is addressed by different scheduling strategies.

C. *Taxonomy*

In this work we cluster the scheduling strategies into two major classes, namely: content-aware and content-unaware strategies. The former is divided into three subclasses and the latter is divided into two broad subclasses, i.e., QoS-aware and QoS-unaware strategies, as shown in Figure 1. Scheduling algorithms that consider the content of the video streams and achieve video optimization using different solutions are categorized under the content-aware strategies. In contrast, scheduling algorithms that consider QoS aspects (e.g., target delay, target throughput and packet error rate) are categorized under the content-unaware strategies. Further details on each of the proposed subclasses are elaborated as follows.

1) First class: content-aware strategies: The radio resource management and packet scheduling solutions for content-aware strategies broadly fall into three subclasses (as shown in Figure 1):

i) *Quality driven scheduling 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 RAN. This type of schedulers consider in the scheduling decision different objective functions (e.g., MSE, PSNR, and 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, as proposed in [15], [17], [28–70].

ii) *Proxy driven radio resource allocation approach* utilizes a simple packet scheduler at the RAN and reduces the congestion by performing video optimization at the cross-layer module either at the video server or at the core network. In this approach, video optimization through cross-layer rate adaptation module at the core network avoids overloading the scheduler at the RAN, as proposed in [71–89].

iii) *Client driven approach* is a video client driven approach, utilizing for instance Dynamic Adaptive Streaming over HTTP (DASH), as proposed in [90–97].

We also, in [98], analyze and compare some of the well known scheduling rules for video streaming traffic. The work involves evaluation analysis in terms of network centric performance metrics as well as user centric metrics, in order to differentiate among the different types of scheduling classes. The evaluation metric of the video quality considers the computation of objective and subjective video quality metrics, whereas the network centric evaluation considers network performance parameters (such as packet loss, system throughput, etc.). The results show that the proxy based video quality aware scheduling strategy performs best in terms of number of satisfied users in LTE networks, and offers high quality video streaming services.

2) Second class: content-unaware strategies: The radio resource management and packet scheduling solutions for content-unaware strategies broadly fall into two subclasses: QoS-aware and QoS-unaware strategies. QoS-aware strategy is represented in the scenario in Figure 2. According to the figure, multi-traffic classes (e.g., Video, VoIP, Best effort) are transmitted from their respective servers to the Evolved Packet Core (EPC). The EPC establishes a connection with an eNodeB via the S1 interface. The MAC layer in the LTE protocol stack is responsible for managing packets and controlling the assignment of the physical resources among flows (i.e., RT and NRT). Examples of the RT and NRT classes are video conferencing, video gaming, and Voice over IP (VoIP) services for the RT class and best-effort services (such as web browsing, FTP, email, etc.) for the NRT class. QoS-aware strategies take into account two important blocks of information, including QoS Class Identifier (QCI) requirements and buffer related information as shown in Figure 2. The QCI is a mechanism in 3GPP LTE systems which ensures that traffic flows are allocated appropriate QoS in terms of latency and packet loss rate. Each traffic flow is assigned a QCI in the EPC. The QCI is represented by a scalar, 8-bit header field, that defines node specific packet forwarding treatment, for instance admission thresholds and scheduling weights at the eNodeB. Different traffic types require different QoS and therefore different QCI values. Examples of the QCI parameters include the class of the flow in terms of guaranteed bit rate (GBR) or Non-GBR, priority value for admission control, packet delay budget, and packet error loss rate. For instance, a conversational voice service flow carries the following QoS parameters: a QCI (1), radio resource type

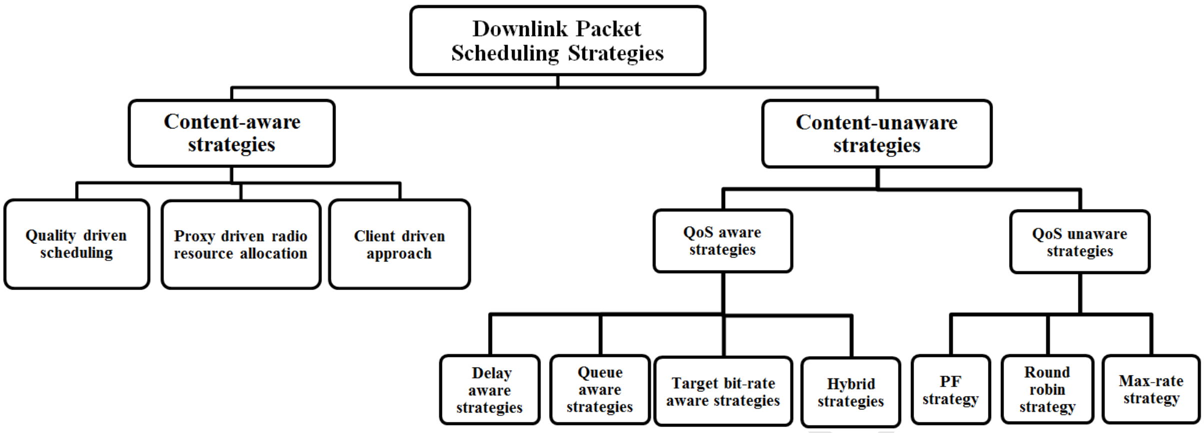


Fig. 1: Downlink packet scheduling approaches classification.

(GBR), priority (2), packet delay budget (100ms), packet error loss rate (10^{-2}).

The second block of information includes the extraction of queue size, Head of Line (HoL) delay, also known as packet latency, and service rate from the buffer at the MAC layer as shown in Figure 2. These information, along with the QCI requirements, are fed to the scheduler in order to define scheduling weights among the flows. Several schedulers (e.g. delay-aware strategy and packet loss fair strategy) are tailored to support various QoS requirements. These QoS requirements must be satisfied by LTE systems by granting the admitted users the maximal balance of fairness and utilization of the service. For instance, delay aware strategy utilizes HoL delay, extracted from the buffer, and target delay information in the scheduling metric to determine the resource allocation priority of different flows. Similarly other strategies extract relevant information from the buffer which depends upon the goal of the scheduling rule. The objective of the scheduling rules varies from meeting end users' flow requirements in terms of packet loss rate, packet delivery delay bounds, target throughput or a combination of these.

i) **QoS-aware strategies**:
We divide the QoS-aware strategies into four categories based on various system and application parameters considered for the scheduling decision (as shown in Figure 1).

a) *Delay aware approach* includes the strategies that utilize the delay parameters in particular, and parameters controlling fairness in general, i.e., Channel State Information (CSI) and average data-rate, in order to perform packet scheduling and radio resource allocations for different users in LTE networks, as proposed in [27], [99–102]. The targets of these schedulers are to increase the number of transmitted packets with a satisfactory level that meets the QoS requirements, have a better delay performance (decrease the probability of violating the target delay of the transmitted packets), provide fairness, and improve the system capacity and efficiency.

b) *Queue aware approach* includes the strategies that utilize the queue size parameter in particular, and parameters controlling fairness in general, in order to perform radio resource allocations and packet scheduling for different users in the wireless networks, as proposed in [103], [104]. The goal of these schedulers is to reduce latency and resource allocation starvation, produce optimal system performance for RT services, and increase the priority of RT flows w.r.t NRT ones.

c) *Target bit-rate aware approach* includes the strategies that are aware of the bit-rate of the flows in the scheduling buffers. The scheduling is performed here by considering GBR and non-guaranteed bit rate (non-GBR) classes for the flows, as proposed in [105–107]. The goal of these schedulers is to maximize the total system throughput, and guarantee the maximum/minimum target bit-rate for a mixture of RT and NRT flows.

d) *Hybrid approach* comprises strategies that utilize a combination of the previous three categories such as queue size, target delay, HoL packet delay, packet error rate and target bit-rate parameters, as proposed in [108–115]. The goal of these schedulers is to maintain the balance and satisfy the QoS requirements for both RT and NRT flows, maintain acceptable MOS level, and increase system capacity and fairness.

ii) **QoS-unaware strategies** include the strategies that utilise parameters controlling fairness, i.e., CSI, and average data-rate, in order to perform radio resource allocations and packet scheduling for different users in the wireless networks. For instance, PF rule, round-robin rule, and max-rate rule are incorporated in this category, as discussed in [116], [117] and shown in Figure 1. The max-rate rule is only aware of the instant bit-rate of the users according to their instantaneous channel conditions (i.e., unlike the target bit-rate in QoS-aware strategies as it considers the rate of the flows). The goal of the aforementioned schedulers in this category is as follows: 1) max-rate rule maximizes throughput but does not satisfy fairness among users, 2) Round Robin satisfies fairness among users but does not maximize throughput,

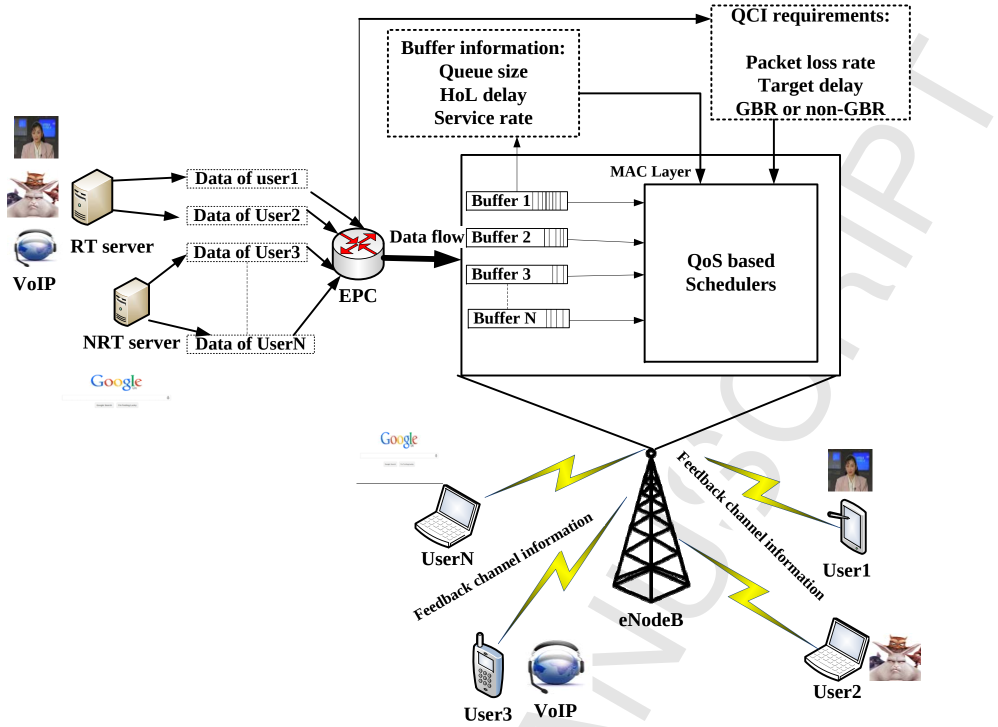


Fig. 2: QoS based scheduling strategies for multi-class traffic.

and 3) PF improves both parameters, i.e., maximizes throughput and maintains fairness among the flows.

Table I summarizes the important approaches and common parameters of the subclasses under the content-unaware class, respectively. The remaining of this survey will focus on content-aware approaches.

III. Content-Aware Scheduling Strategies

This section compares and contrasts the content aware scheduling strategies.

In this section, the MAC layer scheduler at the wireless access network considers the video quality using a utility function. With content-aware scheduling approaches, the optimization goal is to maximize the video quality level subject to time varying wireless capacity. For instance, Figure 3 shows a typical content-aware scheduling scenario where the scheduler is responsible for optimally sharing the radio resources among video flows with diverse video contents and bit-rate requirements.

Different types of content-related information can be considered for the scheduling decision. For instance, video content complexity can be taken into account. Variation is measured, for instance, by calculating the spatial index (SI) and temporal index (TI) [119] which will then identify the complexity level of a video sequence. As an example, video content can be classified into three levels of complexity: low complexity, medium complexity, and high complexity. Complexity is defined here according to the variation in/among the frame(s) and the different bit-rates of different video sequences (i.e., the higher the complexity the higher the bit-rate).

Based on the modality, content/video quality related information is signaled / used by the scheduler, we have identified three classes for scheduling strategies for video optimization over wireless networks. Their detailed description is reported below.

A. *Quality driven scheduling*

In this class, the scheduling approaches are divided into two categories: 1) objective video quality based scheduling utilizing non-scalable video and 2) objective video quality based scheduling utilizing scalable video.

1) Objective video quality based scheduling by utilizing non-scalable video: This section addresses scheduling strategies where a utility function based on objective video quality is adopted and non scalable video is considered. Most of these strategies consider a measurement of the distortion associated to the loss of specific packets.

In [34] and [35], the concept of incrementally additive distortion is used to determine the importance of video packets for each user for pre-encoded non-scalable video streams. 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 lowest importance video packet. The authors proposed dropping strategies based on the packet's importance towards

TABLE I: Channel and QoS Parameters Used by State-of-The-Art Content Unaware Downlink Scheduling Approaches.

Scheduling Approaches	**Channel Parameters**		**QoS Parameters**					
	CSI	*Average data rate*	*PLR*	*Target rate*	*Queue size*	*Service aware*	*Target delay*	*HoL delay*
Delay aware approaches								
M-LWDF [118]	X	X				RT	X	X
EXP/PF [99]	X	X				RT, NRT	X	X
DAPS various traffic classes [101]	X					RT, NRT	X	X
EXP-rule [102]	X	X					X	X
Log-rule [102]	X	X					X	X
Queue aware approaches								
PFPS [103]	X	X			X			
VT-M-LWDF [104]	X	X			X	RT, NRT		
Target bit-rate aware approaches								
Target bit-rate rule [105]	X	X		X				
Target bit-rate rule [106]	X			X				
Target bit-rate rule [107]	X			X				
Hybrid approaches								
Queue-HoL-M-LWDF [108]	X	X			X	RT, NRT	X	X
Modified-PF [109]	X			X		RT, NRT	X	
Joint Channel and Queue [110]	X	X			X	RT, NRT	X	X
Hybrid approach [111]	X	X	X	X		RT, NRT	X	
DPS [112]			X	X		RT, NRT	X	
FLS [113]	X	X			X	RT, NRT	X	
QoS-unaware approaches								
PF [116]	X	X						
Maximum-rate [117]	X							
Round Robin [117]								

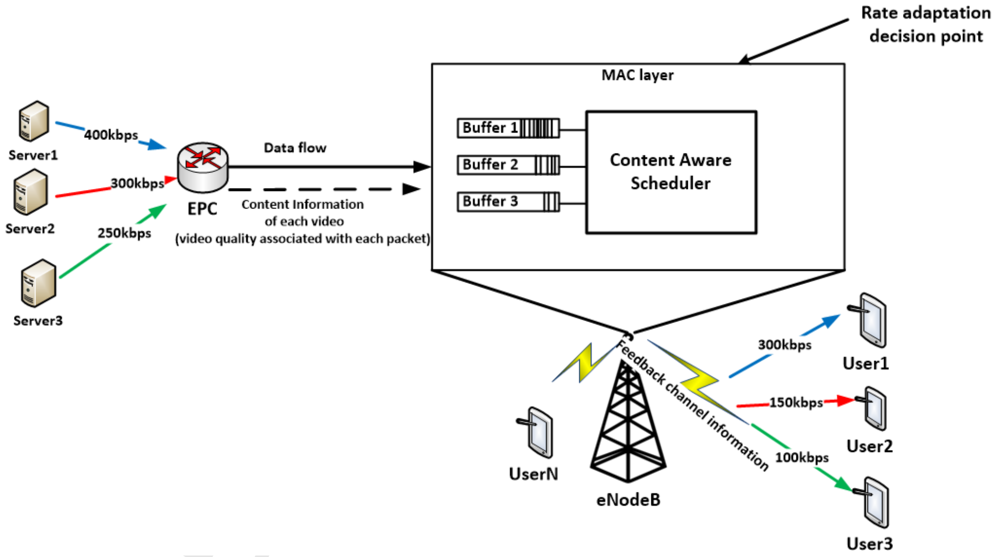


Fig. 3: Objective video quality based scheduling.

the video quality as well as packet's waiting time in the queue and combined these dropping strategies with state-of-the-art packet scheduling rules (PF, M-LWDF and Max throughput). The proposed dropping strategies at the entrance of the wireless access system have a similar impact as server-based rate control schemes where the congestion signal is sent to the server by the wireless access network. Although the impact on distortion of the loss of a specific packet depends on its content, their proposed strategies consider the content of the video packets only in terms of dependency from other packets.

Moreover, the approach in [66] focuses on the development of QoE-aware optimization downlink scheduling video traffic flow. The aim of this work is to maximize QoE of video traffic streaming over LTE networks. A novel integration framework between genetic algorithm (GA) and Random Neural Network (RNN) is applied to QoE-aware optimization of video stream downlink scheduling. The proposed framework was compared with other state-of-the-art LTE downlink scheduling algorithms such as FLS, EXP-rule, and LOG-rule. Under different network conditions, simulation results showed that the proposed scheduler can achieve better performance in terms of QoE (10% increase), throughput and fairness.

A quality-driven resource allocation and scheduling for delay-constrained video transmission (which has different tolerable delay bounds such as live streaming or two-way video conferencing) is proposed in [68]. The approach relies on a multiuser setup where different users have different delay QoS constraints. The derived resource allocation strategy is designed to maximize the sum video quality which is measured through concave rate-quality mapping with respect to any quality metric. Furthermore, the proposed approach solves the problem related to the fairness of resource allocation by maximizing the minimum video quality across users. The goal of scheduling approach in this work is to enable selecting a maximal user subset such that all selected users in this subset can meet their statistical delay requirements (i.e., their QoS requirement). Based on the provided results, significant gains in capacity, measured in terms of number of video users supported in the system, are achieved due to QoS-aware scheduling and resource allocation.

Furthermore, a QoE-aware video adaptation and resource allocation approach for power-efficient streaming over downlink OFDMA systems is proposed in [69]. The adaptation scheme selectively drops packets from a video stream to produce a lower bit-rate version which results in a reduction of delay and a satisfaction of a target users QoE. The resource allocation target is to minimize the transmission power by considering the delay limitation of the lower rate stream. Experimental results show significant performance enhancement of the proposed system in terms of reducing end-to-end delay and power consumption while satisfying QoE requirements.

In multimedia applications, the content of a video packet is critical for determining the contribution of the packet to image/video quality. The content-dependent utility gained due to packet transmission can be considered in the scheduling rule in addition to other information such as the size of the packet in bits and the decoding deadline for the packet, i.e., each frame's time stamp. The authors in [36] and [37] investigate a content-aware resource allocation and packet scheduling for video transmission over wireless networks. The authors present a cross-layer packet scheduling approach, transmitting pre-encoded video sequences over wireless networks to multiple users. This approach is used for Code Division Multiple Access (CDMA) systems and it can be adapted for OFDMA systems such as IEEE 802.16 and LTE wireless networks. The data-rates of the served users are dynamically adjusted depending on the channel quality and the gradient of a content-aware utility function, where the utility function takes into account the distortion of the received video. The method adopted in [36] and [37] consists of ordering the packets of the encoded video according to their relative contribution to the final quality of the video, and then constructing for each packet a utility function whose gradient reflects the contribution of the packet to the perceived video quality. Hence, the utility function is defined as a function of the decoded video quality, based on the number of packets already transmitted to a user for every video frame. Further, robust data packetization at the encoder and realistic error concealment at the decoder are considered. The proposed utility function enables optimization in terms of actual quality of the received video. The authors provide an optimal solution when video packets are decoded independently and a simple error concealment approach is used at the decoder. Moreover, with complex error concealment a solution is provided where a distortion utility function is calculated. Performance evaluation is carried out and it is noticed that the proposed content-aware scheduler outperforms content-independent approaches in particular for video streaming applications. The parameters used in this scheduler are the achievable rate, CSI from UEs, weighting parameter for fairness purposes across users (which is based on the distortion in a user's decoded video in the previous transmissions) and three features of each packet: utility gained due to packet transmission, decoding deadline, and packet size.

The authors in [39] further extend the concept of video packet importance and propose a joint packet scheduling and subcarrier assignment for OFDMA systems. According to [39], the subcarrier with the largest contribution for the overall video quality of users' is allocated through a two-level search path. First, for each flow the video packet contributing largest to the video quality is selected. In the second step, the priority of each flow is computed by considering the channel gain of the subcarrier and the importance of packet. In other words, the subcarrier is assigned to a flow having the best channel gain and the video packet contributing largest to the video quality. A similar approach is presented in [40] where a flow is prioritized based on the ratio of the packet's contribution to video quality 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 lowest number of subcarriers is scheduled in priority. Similar distortion based joint packet scheduling and subcarrier assignment policies are presented in [41–43], [62]. The main goal of these strategies is to maintain fairness across different users and minimize the received video distortion of every user by adopting a fine granular video quality based packet scheduling and radio resource allocation approaches. It is important to note that strategies in [39–43] drop the video packets violating the pre-assigned delay bound. However, none of the strategies above consider packet deadline in the scheduling decisions.

The work in [60] proposes a QoE-aware resource allocation scheme for High Efficiency Video Coding (HEVC) encoded video transmission over LTE networks. In this approach, the contents of the video are prioritized based on the importance of the slices of the compressed video. The importance is considered based on the actual contributions to the motion compensation. Thus, the important video slices are allocated the most robust resource blocks (RBs) to guarantee successful delivery over error prone channels and to ensure better QoE for the users.

The authors in [44] investigate an application driven cross-layer approach for video transmission over OFDMA based networks. The proposed schemes are named quality-fair, and quality-maximizing. They are used to maximize the video quality with and without fairness constraint, respectively. The packet scheduling will be responsible for selecting the packets with the largest contribution to the video quality. The assumption in this design is that each video frame is partitioned into one or more slices, with each slice header acting as re-synchronization marker to allow independent decoding of the slices, and where each slice contains an integer number of macro-blocks. Hence, due to the variation of the contents among different video streams, different packets make diverse contribution to the video quality. The parameters used are a quality contribution index for each packet, expressed by the decreased distortion value caused by the successful transmission of the packets, size of the packets, maximum delay, real time requirements for video applications, and the CSI fed-back by the mobile station.

In [45] the authors presented a scheduling algorithm for an OFDMA system that can be tuned to maximize the throughput of the most significant video packets, while minimizing the capacity penalty due to quality/capacity trade-off. It is shown that the level of content-awareness required for optimum performance at the scheduler and the achieved capacity are highly sensitive to the delay constraint. The tighter the delay constraints, the higher the importance of content-awareness and lower the number of satisfied streaming users. The study shows that the consideration of delay awareness in the scheduling decision enhances the video quality.

The authors in [46] propose a novel cross-layer scheme for video transmission over LTE wireless networks. This scheme takes information from the three layers of the Open Systems Interconnection (OSI) model namely: Application (APP), MAC, and Physical (PHY) layers. The I-based and P-based packets are extracted from the APP layer, the packets are scheduled and prioritized at the MAC layer and the channel state information is taken from the PHY layer. The work in [46] aims at improving the perceived video quality for each user and improving the overall system performance in terms of spectral efficiency. It is assumed that *I* packets are more important than *P* packets for each user. For the reason that is because the loss of an I packet may lead to error propagation within the Group of Pictures (GOP). Hence, the packet scheduling algorithm at the MAC layer is adapted to prioritize *I* packets against *P* packets for each video sequence. Results show that the proposed cross-layer scheme performs better in terms of system throughput and perceived video quality. The parameters used are achievable rate, service rate requirements of *I* and *P* packet queues, and CSI.

The authors in [47] propose a distortion-aware scheduling approach for video transmission over OFDMA based LTE networks. The main goal of this work is to reduce the end-to-end distortion in the application layer for every user in order to improve the video quality. Hence, parameters from the PHY, MAC and APP layers are taken into consideration. At the APP layer the video coding rate is extracted, at the MAC layer PRB scheduling and channel feedback are exchanged, and at the PHY layer modulation and coding parameters are used. Parameters used are frame distortion caused by lost slices, waiting time, transmitting time, latency bound, video distortion caused by BLER (a function of modulation and coding scheme of PRB and SINR of wireless channel), the dependency of the video content under the constraint of the transmitting delay, and different coding rates. Simulation results show that the proposed gradient-based cross-layer optimization can improve the video quality.

The authors in [63] propose a cross-layer design scheme for optimizing resource allocation for H.264 video applications over LTE networks based on QoE evaluation. The work involves three steps: 1) a MOS-based QoE prediction function is introduced to maximize the perceived video quality of users while maintaining fairness among them, 2) a mapping model, which reflects the relationship between the objective parameters and the subjective perceived video quality (i.e., PSNR and MOS), is proposed, and 3) Particle Swarm Optimization (PSO) is utilized to explore the optimal strategy for users according to their QoE prediction values. The goal of the proposed approach is to improve the perceived video quality as well as to maintain the fairness between users when compared to the traditional scheduling schemes.

A QoE-aware scheduler aiming at maximizing the minimum MOS in a system subject to attaining at least a given number of satisfied users is presented in [67]. The proposed approach takes into account all the data that the UEs have already received, and not only the data of a specific instant of time, as well as it excludes from scheduling process the UEs that are already satisfied. According to simulation results, the proposed solution not only improves the users' satisfaction and the minimum experienced MOS in the system, but it also doubles the capacity in terms of supported number of users compared to benchmark solutions. It is also able to handle imperfections on the estimation of the CSI.

The authors in [64] introduce a cross-layer QoE-aware downlink radio resource allocation approach for OFDMA systems. The aim of the proposed algorithm is to ensure an appropriate level of QoE for each user via considering application-layer parameters and human perceived video quality into the radio resource allocation process. The work comprises a design of two algorithms: the first algorithm dynamically allocates resources by guaranteeing the same QoE level/maximizing the minimum experienced MOS to all users through a utility function. Whereas the second algorithm presents a reliable balance among the user's QoE level and the system spectral efficiency. These two algorithms are considered user-centric approaches, as they assure the user's QoE. The consideration of the application-layer parameters (such as video resolution, GOP encoding structure, and PSNR) and user's video perceived quality into the resource allocation process leads to achieve high users' QoE and data-rate control. The algorithms achieved significant increase in the QoE level and a fair distribution of resources among users.

2) Objective video quality based scheduling by utilizing scalable video: Scalable Video Coding (SVC) [120] represents a video sequence via multiple layers with different quality, resolution, and/or frame rate as shown in Figure 4. SVC enables graceful degradation of video quality when resources are limited, hence it is particularly suitable for the case of multi-user video scheduling.

A content-aware downlink packet scheduling scheme for multi-user scalable video delivery over wireless networks is proposed in [48]. The scheduler uses a gradient-based scheduling framework, as elaborated earlier in [37], along with scalable video coding schemes. The reason for using

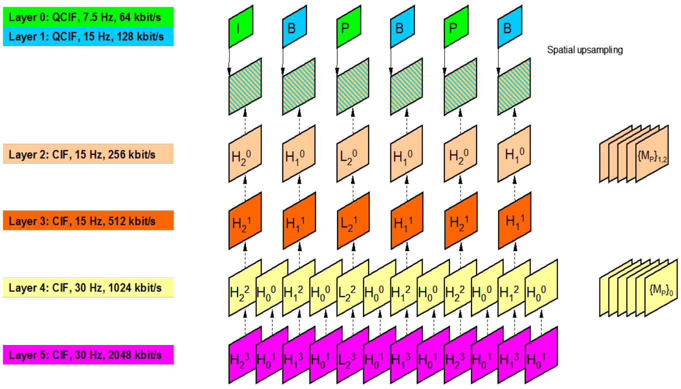

Layer 0: QCIF, 7.5 Hz, 64 kbit/s
Layer 1: QCIF, 15 Hz, 128 kbit/s
Layer 2: CIF, 15 Hz, 256 kbit/s
Layer 3: CIF, 15 Hz, 512 kbit/s
Layer 4: CIF, 30 Hz, 1024 kbit/s
Layer 5: CIF, 30 Hz, 2048 kbit/s
I
B
P
B
P
B
Spatial upsampling
H_2^0
H_1^0
L_2^0
H_1^0
H_2^0
H_1^0
$\{M_P\}_{1,2}$
H_2^1
H_1^1
L_2^1
H_1^1
H_2^1
H_1^1
H_2^2 H_0^0 H_1^2 H_0^0 L_2^2 H_0^0 H_1^2 H_0^0 H_2^2 H_0^0 H_1^2 H_0^0
$\{M_P\}_0$
H_2^3 H_0^1 H_1^3 H_0^1 L_2^3 H_0^1 H_1^3 H_0^1 H_2^3 H_0^1 H_1^3 H_0^1

Fig. 4: Temporal, Spatial and Quality scalability of SVC.

SVC is to provide multiple high quality video streams over different prevailing channel conditions for multiple users. The scheduler proposed in [48] outperforms the traditional content-independent scheduling approaches. Packet prioritization can be signaled to the MAC layer scheduler, in conjunction with the utility functions of each packet. A distortion model is also proposed in order to efficiently and accurately predict the distortion of an SVC encoded video stream. The model is employed to prioritize source packets in the queue based on their estimated impact on the overall video quality. The parameters used are achievable rate for each user, loss probability, user's estimated channel state, and expected distortion.

The authors in [49] and [50] propose a scheduling and resource allocation strategy for multi-user video streaming over OFDM downlink systems. The authors utilize SVC for encoding the video streams. This work utilizes only the temporal and quality scalability (not spatial scalability) for video coding via the adaptive resource allocation and packet scheduling. The authors propose a gradient-based packet scheduling and resource allocation strategy. This strategy prioritizes different users by considering adaptively adjusted priority weights, computed based on the video content, deadline requirements and transmission history. A delay function is designed in order to cope with the effect of the approaching deadline, for which the possibility of delay violation is reduced. The aim of the work presented in [49] and [50] is to maximize the average PSNR of all SVC video users under a constrained power transmission, time varying channel conditions, variable rate video contents, and video frame deadline. The obtained results show that the proposed scheduler outperforms the content-blind and deadline-blind algorithms with a gain of as much as 6 dB in terms of average PSNR when the network is congested.

A quality-aware fair downlink packet scheduling algorithm for scalable video transmission over LTE systems is proposed in [51]. The authors proposed a Nash bargaining based fair downlink scheduling strategy for scalable video transmission to multiple users. A novel utility metric based on video frame importance in a 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 quality-based utility in comparison with advanced, state-of-the-art throughput based strategies. The authors in [121] improve the work in [51] by exploiting multi-user time-averaged diversity.

The authors in [65] propose a QoE-aware video streaming solution to maximize multiuser QoE for Scalable Video Coding (SVC) streaming over multiuser (MU) Multiple-Input Multiple-Output (MIMO) Orthogonal Frequency Division Multiplexing (OFDM) systems. The proposed approach is achieved by combining QoE-aware video adaptation (QoEVA) and QoE-aware resource allocation (QoERA) schemes. The QoEVA of SVC is analyzed via a subjective video quality assessment database to derive QoE-optimized scalability adaptation tracks. A rate-QoE model is then developed to approximate the track and is utilized to design a QoE-based resource allocation scheme. The enhancement of multiuser QoE is proven in this study, where experimental results show that the proposed QoEVA significantly performs better than the conventional video adaptation schemes and the proposed QoERA achieves much better user experience when compared to state-of-the-art solutions.

A channel and content aware 3D video downlink scheduling combined with a prioritized queuing mechanism for OFDMA systems is proposed in [52]. The idea behind the queuing mechanism is to prioritize the most important video layers/components with the goal of enhancing the perceived 3D video quality at the receiver. The work focused on color plus depth 3D video and considered the unequal importance of different components with respect to the perceived quality. 3D video is encoded using an SVC video encoder. The priority values of the encoded 3D video components are signaled from the application layer to the MAC layer via cross-layer signaling. The users feedback to the base station their sub-channel gain, which is then used at the MAC layer for the resource allocation process. Hence, the proposed scheduler is designed to guarantee that the important layers are scheduled at every scheduling epoch over the sub-channels with higher gain. The Packet Loss Ratio (PLR) is decreased for the prioritized color/depth layers at the MAC layer at the expense of a little increase in the PLR for the less perceptually important video layers. Video layers expected to be highly affected by packet losses are discarded to avoid wastage of radio resources. The prioritization scheme results in global quality improvement. The parameters used are the HoL packet delay, a weight that controls the throughput fairness among users, the fractional rate based on video layer bit-rate, SINR, dependency/temporal/quality (DTQ) identifications of the SVC video stream, and the maximum tolerable delay based on the required playout time, i.e., based on the frame rate.

The aforementioned approaches require complex application layer information, such as distortion (if a 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. These algorithms require extensive cross-layer signaling as well as Deep Packet Inspection (DPI) is a procedure used to identify the importance of a packet towards the overall video quality) so that all the associated information can be conveyed to the scheduler at the eNodeB. It is important to note that different requirements of video content information increase the scheduling complexity at the MAC layer of the base

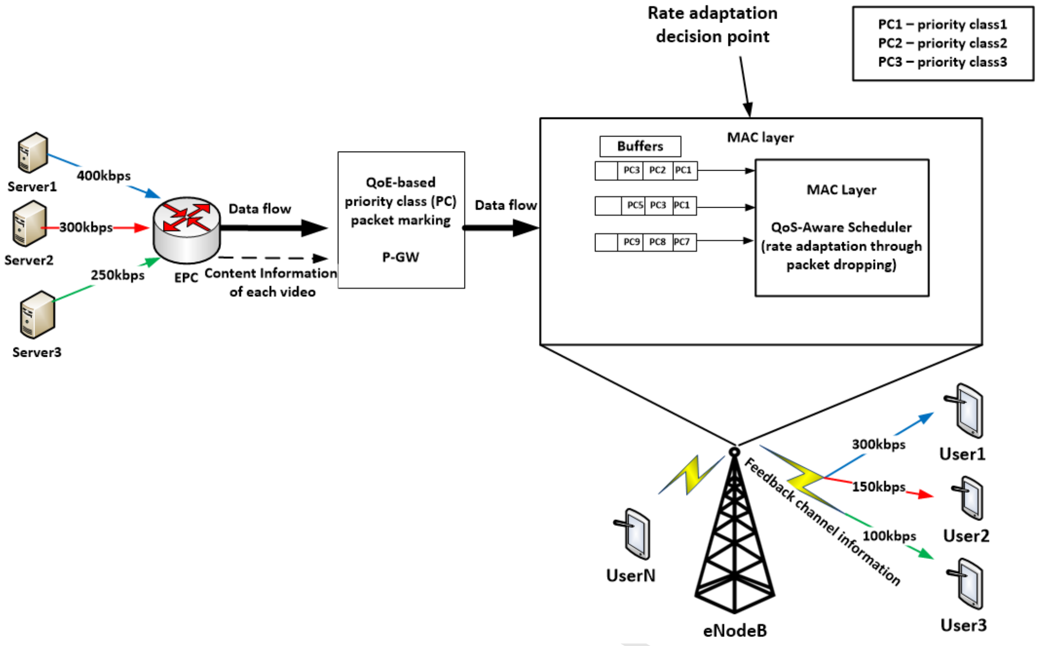


Fig. 5: QoE-aware packet priority marking based scheduling.

station. Therefore, such scheduling algorithms pose problems from an implementation point of view. The authors in [53] discuss a cross-layer system design between the streaming server and the mobile base station for SVC streaming. The authors propose to map SVC layers to QoS service classes. According to their study, a considerable quality gain is achieved when a base layer is assigned the most important service class and enhancement layers are assigned to the least important service class. Such mapping strategies for SVC can provide graceful quality degradation for a single video by utilizing DPI such as in [54], [55]. However, they do not allow to compare and prioritize layers of multiple videos having different quality and rate characteristics. The work done in [56], [59], [122] provides a QoE-based packet marking avoiding DPI at the RAN. Furthermore, the packet marking strategy in [56], [59], [122] allows network operators to adapt multiple video streams having different video layers and diverse quality and rate characteristics.

Figure 5 illustrates the basic idea of the QoE-based packet marking strategy. 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 overall QoE 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 [123]. This class of content-aware scheduling avoids the extensive amount of information (distortion, decoding dependence, etc.) transfer to the eNodeB as only Priority Class Index (PCI) information is required at the eNodeB. According to [56], [59], QoE-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.

B. Proxy driven radio resource allocation

The aforementioned approaches perform cross-layer optimization with the goal of performing efficient link layer packet scheduling for delay sensitive video traffic. The scheduler specifically considers the video content with the goal of maximizing the number of satisfied users. In proxy driven approaches, the main idea is to gather key parameters across the application, MAC, and physical layers in a decision center (optimizer). In some cases more than one decision centers are considered. The cross-layer optimization framework allows the network operators to perform radio resource allocation based on users' satisfaction. For solving wireless multimedia radio resource allocation problems, there is a good amount of research work being carried out using cross-layer optimization techniques. According to [71–74], the main steps involved in a proxy driven cross-layer optimization framework are:

- Collecting key information from each of the layers through cross-layer signaling;
- The cross-layer information is gathered in a decision center (also called video optimizer as shown in Figure 6). The decision center performs the overall optimization by considering the variables from each of the layers like

channel quality, time-averaged throughput, video content information, and buffer status;

- After the joint optimization, the decision center sends the resource allocation information to the MAC layer. The scheduler then performs the packet scheduling based on the radio resource allocation decisions. If any rate adaptation is required then the optimizer performs the rate adaptation either through transcoding or packet dropping as shown in Figure 6.

1) Proxy driven radio resource allocation utilizing non-scalable video: As we mentioned earlier, the previous cross-layer optimization techniques only consider the video application and design a scheduling rule by considering the video content and channel quality. The authors in [75] propose a general resource allocation framework which can accommodate all the traffic classes (VoIP, video, web browsing, FTP). The authors propose MOS as the optimization metric for each of the traffic types and design a utility metric which quantifies the users' satisfaction in terms of the MOS. In [76] the authors further extend the work by introducing different objective functions such as the modified max-min MOS where the objective function guarantees a minimum MOS to each of the flows. The cross-layer optimization technique by considering the MOS of each application shows a remarkable improvement in terms of the number of satisfied users as compared to the throughput-based resource allocation schemes. The cross-layer optimization technique considered in [75], [76] assumes that the video server is located close to the base station which allows quick rate adaptation of the video application. The rate adaptation is assumed to be performed by changing the quantization parameters of the video streams. However, video servers are in general located outside the wireless RAN thus limiting flexibility of delay-free rate adaptation. Cross-layer optimization techniques involve extensive cross-layer signaling which increases the overheads and involves additional delay. In dynamic wireless environment, timely rate adaptation is very important. When the video server is located outside the RAN, the additional delay imposed in the end-to-end link probing from video server to the base station decreases the probability of timely rate adaptation. Thus, the performance of the cross-layer optimization strategy decreases under congestion. Furthermore, the cross-layer optimization strategy requires the video server to support the required cross-layer signaling protocols such as [77–79] to support the video rate adaptation thus adding compatibility issues.

The European projects PHOENIX [80], [81] and OPTIMIX [82] proposed a framework where the cross-layer adaptation task is split in two main control entities. The application layer ("Application controller") performs rate-control and adaptation based on information from the lower layers, whereas the PHY/MAC layer ("Base station controller") adapts PHY/MAC parameters based on the characteristics of the video flows. The two controllers are supposed to work on different timescales and require information obtained through cross-layer signaling.

The works in [83–86] propose a realistic scenario in which the video streaming content is stored at the video server outside the RAN. In order to perform "in network rate adaptation" they propose two modules which are 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. With the two modules located in the RAN, the proposed resource allocation optimization cycle is 1 sec. The Traffic Engineering module performs rate adaptation either based on packet dropping or transcoding. The authors in [85] 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, the resources are first reserved to the users with good channel quality and applications demanding low data-rates. 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 (the minimum quality across the users is maximized). 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. In order to reduce the resource allocation optimization cycle to 1 sec, the framework proposed by [83–86] requires extensive cross-layer signaling and substantial investments on new modules which can reduce the congestion through fast rate adaptation of video traffic at the RAN.

2) Proxy driven radio resource allocation utilizing scalable video: The work done in [87] jointly addresses resource allocation and rate adaptation for SVC traffic. The authors propose 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 authors consider three key elements, i.e., the Multimedia Provider (MP), Media Aware Network Element (MANE) and the OFDMA-based Wireless Access Network (WAN). The mobile users in the WAN request the video streaming from the MP. The MP streams the video sequences encoded according to the SVC standard with temporal and quality scalability. The MP sends the R-D (rate-distortion) information of the video sequences as well as the priority index of each video frame to the MANE. The MANE also receives the buffer status and link channel quality of the streaming users and performs resource allocation for each of the video streaming flows. The authors proposed an optimal solution based on iterative local approximation (ILA). With the ILA algorithm, the information exchange between the application (MANE) and MAC (WAN) layer is substantially reduced. The ILA algorithm at MANE is executed at regular time intervals (in the order of seconds). The results show a PSNR gain of 7 dB for high complexity videos as compared to the algorithms having similar complexity. However, the ILA algorithm considers a strong assumption that the channel quality of the mobile users remains the same during each

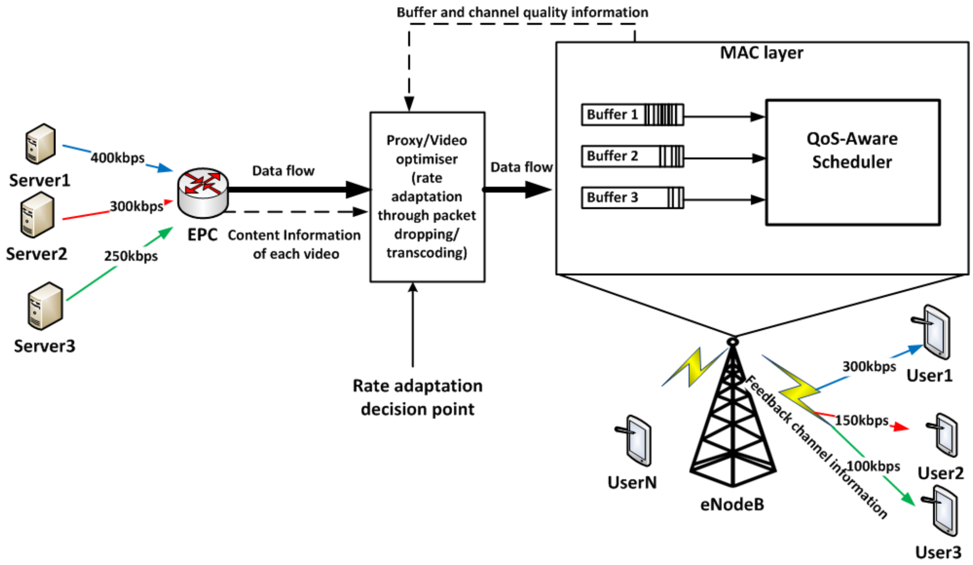

Buffer and channel quality information
MAC layer
Buffer 1
Buffer 2
Buffer 3
QoS-Aware Scheduler
Proxy/Video optimiser (rate adaptation through packet dropping/ transcoding)
Data flow
Data flow
400kbps
300kbps
250kbps
Server1
Server2
Server3
EPC
Content Information of each video
Rate adaptation decision point
300kbps
150kbps
100kbps
User1
User2
User3
UserN
eNodeB
Feedback channel information

Fig. 6: Proxy driven radio resource allocation.

optimization cycle which spans over a few seconds.

The authors in [89] presents a study on optimal caching strategy of SVC streaming in small cell networks with respect to channel diversity and video scalability. The QoE-aware proactive caching of scalable videos approach is formulated through an integer programming problem to maximize the average SVC QoE. This is achieved by establishing connections between QoE and caching state of each video. Simulation results show that the small cell base stations with caching ability can greatly improve the average QoE of SVC streaming service, and that the proposed caching strategy acquires significant performance gain compared with other conventional caching strategies.

The authors in [88] propose a content-aware scheduler to allocate resources for downlink scalable video transmission over LTE based systems. The goal of the scheduler is to maximize the average video quality across all users. The scheduler calculates, on a frame by frame basis, the proportion of PRBs required by each of the video streaming users subject to the constraints of channel quality, total number of available PRBs and device capability. In LTE, one frame corresponds to 10 TTIs and 1 TTI is equal to 1 ms. Therefore, the content-aware scheduler determines the number of PRBs required by each of the video streaming user, whereas the specific PRBs are finally allocated by a TTI level schedulers discussed in Section II-C2. The authors also propose a detailed signaling mechanism between the content aware PRB scheduler and the TTI level scheduler. Furthermore, they also propose the required architecture modification for the implementation of the content-aware PRB scheduler. They propose the implementation of the content-aware scheduler either in the Mobility Management Entity (MME) or at the eNodeB. In addition, the study also presents the impact of signaling delay on the video distortion in the indoor and outdoor environments. According to their study, the video quality is substantially reduced in urban environment, where the link quality changes every TTI because of the multi-path fast fading, shadowing and Doppler effects. Under such dynamic environment, the content-aware PRB scheduler receives outdated channel quality information since it operates every 10 TTIs. The parameters used are the SVC profile levels, CQI, number of available PRBs along with a quality-driven scheduler. The video quality is measured using two full-reference methods, i.e., relying on the knowledge of the the original video: PSNR and the SSIM metric.

C. Client driven approach

The aforementioned strategies consider Real-Time Protocol / User Datagram Protocol (RTP/UDP)-based streaming which can often be blocked by firewalls. Recently, Dynamic Adaptive Streaming over HTTP protocol (MPEG-DASH) has been standardized for mobile multimedia applications. The standard comprises a conventional approach of HTTP/TCP over IP networking. In DASH-based streaming, video is divided into chunks (also known as segments) with each chunk encoded into multiple bit rates *i.e.*, DASH representations. During the course of streaming, a client can adapt the video streaming bit rate by requesting the subsequent chunk based on the available network capacity, as shown in Figure 7. Therefore, video rate adaptation decisions are taken by the clients. Video server reports the available bit-rate of all the dash representations of each segment through a Media Description File (MDP). In [90], authors analyze the rate adaptation strategies of three well known commercially available DASH clients. According to their experiments, all the rate adaptation algorithms have a high response time under congestion, i.e., they react very slowly to throughput variations. In mobile video streaming, such rate adaptation strategies are not useful because of the probabilistic arrival of the incoming traffic and stochastic nature of the wireless channel which leads to high throughput fluctuations. The authors in [91] propose a rate adaptation strategy for mobile clients by considering the TCP throughput, status of the video playout buffer and available battery time.

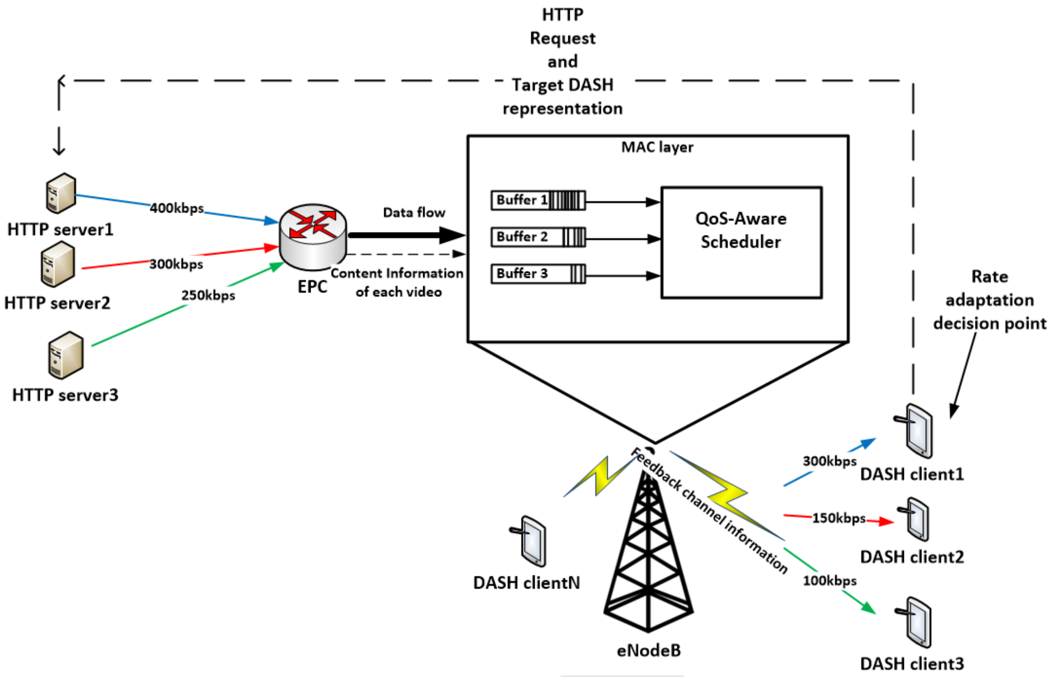


Fig. 7: Client driven approach.

The video rate adaptation algorithm consists of two modules. The first module predicts the target bit rate based on status of the playout buffer (available buffered video time) and achieved TCP throughput. The main goal of this module is to minimize the freezing delay which occurs when the buffer gets exhausted. The second module predicts the target bit rate based on the available battery time. The main goal of the second module is to maximize the available video watching time due to battery constraints. The bit rate of the requested chunk is the minimum of the two target rates predicted by the two modules. The algorithm efficiently absorbs dynamic TCP throughput fluctuations and achieves smooth video streaming in an energy efficient manner. Similarly, the authors in [124] proposed a playout buffer aware scheduling strategy, where client's playout buffer depletion probability is minimized. However, in multiple adaptive HTTP streaming over wireless networks two major issues exist:

- Multiple video streaming clients sharing the same bottleneck result in an unfair bandwidth sharing. For instance, the authors in [90] analyze the fairness and stability (variations in video quality per unit time) performance of two HTTP streaming clients sharing the same bottleneck. According to the analysis, the two streaming clients sharing the same access network result in an unfair distribution of bandwidth. Generally, in HTTP streaming, when a client starts streaming it requests the lowest possible bit rate. It then switches to the highest available bit rate depending upon the achieved throughput. If a new video streaming client joins the network, it stays to the lowest available bit rate because the other client is utilizing the remaining available bandwidth.
- User context, in terms of client's playout buffer level and video content, information is not available at the RAN. Therefore, the client-based rate adaptation decisions are not QoE driven in terms of radio resource optimization for multiple users. In [92], the authors proposed to add rate-quality information to the Media Presentation Description (MPD) so that the client can select the optimal bit-rate, *w.r.t* the video quality. However, this approach maximizes the video quality for a single user without considering the rate-quality characteristics of other users sharing the same resources. In cellular networks, for instance, multiple users are competing for resources and the DASH server lies outside the operator's network. Therefore, operators have control on the radio resource allocation to multiple users but not on the video rate adaptation decisions. This can create congestion at the base station, if HTTP clients react slowly to the available bandwidth variations.

In order to solve the aforementioned issues, the authors in [93] propose a solution to re-write the client request at a proxy close to the eNodeB. They assume that the DASH server adds the rate-quality information of each video segment to the MPD which is then exploited by the QoE optimizer. The QoE optimizer collects the channel quality information of different clients from the eNodeB. The QoE optimizer allocates bandwidth to the video clients based on rate-quality characteristics and the available network resources. Based on QoE optimization results, the HTTP requests from each of the clients are re-written at the proxy. Simulation results show that proactively re-writing the HTTP requests of the DASH clients, at the proxy, shows a remarkable improvement in the video quality as compared to relying on the rate adaptation logic of the DASH clients. However, the authors do not consider the buffer status of the DASH clients in re-writing the HTTP requests. It is important to note that DASH comprises a buffered video streaming and proactively re-writing the HTTP requests without considering the buffered video time may underutilize the network resources.

Furthermore, the authors in [94] propose a QoE aware radio resource allocation for HTTP adaptive streaming over LTE systems. The approach is designed to constrain re-buffering probability for adaptive streaming users. The proposed algorithm is called a Re-buffering Aware Gradient Algorithm (RAGA), which depends on periodic media buffer feedback (i.e., already standardized in the DASH standard) from adaptive streaming clients. The RAGA algorithm shows high reduction in the re-buffering percentage for adaptive streaming users without compromising the video quality. On the contrary, the authors in [95] design a video aware resource management at the network core rather than at the network edge, as this approach enhances video quality of experience in QoS-aware networks like LTE. Further, in their proposed architecture, a Video Aware Controller (VAC) is located at the network core, where it periodically receives HTTP adaptive/video streaming related feedback from adaptive streaming clients/servers. Then, the VAC controller converts the information into QoS parameters for each user, as VAC acts as the central element for managing video aware resources. The QoS parameters are then signaled to multiple network elements as well as HTTP adaptive streaming clients. On the other hand, at the MAC layer, an algorithm is designed to periodically compute the maximum bit-rate for each streaming user/network flow based on the provided media buffer feedback from clients to the VAC controller. This architecture shows further reduction in re-buffering percentage and better perceived video quality.

The authors in [97] propose three methods to improve the QoE of SVC-based DASH users over next generation wireless systems. The methods are: 1) design of an improved mapping scheme from SVC layers to DASH representations that can provide the desired bit-rates, enhance the throughput, and reduce the HTTP communication overhead, 2) development of a DASH-friendly scheduling and resource allocation algorithm by integrating the DASH-based media delivery and the radio-level adaptation via a cross-layer approach. This method utilizes the characteristics of video content and scalable video coding, and greatly reduces the video stalling probability by considering the client playout buffer level, and 3) proposal of a DASH proxy-based bit-rate stabilization algorithm to improve the video playout smoothness that can achieve the desired trade-off between playout quality and stability. Results show that the proposed schemes achieve better performance than existing resource allocation methods.

Table II summarizes the important approaches, techniques and common parameters of the three subclasses under the content-aware class, respectively.

IV. Standardization Efforts for QoE-aware video delivery

The challenges on mobile operators for the implementation of content-aware scheduling strategies are multi-fold. In this section, we highlight key steps taken by 3GPP for QoE enhancements over LTE. In recent years, the 3GPP standardization body has put in efforts to identify issues as well as provide solutions, with the goal of maximizing the QoS and QoE for the end-users [25], [125–127]. In the following, we summarize the solutions proposed by the 3GPP for the provision of QoE-aware video delivery over LTE networks.

A. Video context aware solution

For content-aware scheduling strategies, the availability of video content information at the eNodeB is crucial. Majority of strategies studied in Section III assume this information available at the MAC layer of the eNodeB, either through explicit cross-layer signaling or DPI. However, the current mobile network architectures do not support eNodeB in obtaining video context information for QoE estimation, *i.e.*, video content and playout buffer levels. Nonetheless, UEs reporting video context information to the eNodeB is currently in the focus of standardization. According to 3GPP TR 36.933 [126], UEs report video context information in terms of playout buffer level to the MAC layer of the eNodeB. These reports can be sent periodically or configured to be triggered by an unexpected event (e.g. network congestion). Such information can be used to avoid unnecessary video stalling and improve client's QoE. According to [126], video playout buffer aware scheduling supports 25 UEs, at 1% video stalling probability level, as compared to the non-context aware PF scheduling rule which supports only 20 UEs. Another important step towards standardization is the users context information reporting in terms of the available DASH representations of the video content, discussed in Section IV-D.

B. Operator specific QoS Class Identifiers (QCIs)

3GPP classifies different types of bearer into different classes by assigning different QCIs, represented by an 8-bit field. According to 3GPP TS 23.203 [127], 15 QCI values have been standardized. The performance characteristics of each QCI have been predefined in terms of edge to edge packet forwarding treatment. The standardization of QCIs allows different network operators to ensure minimum QoS levels for different services and applications. The introduction of video content based packet prioritization, as studied in Section III-A2 and shown in Figure 5, requires different packet forwarding treatment for different packets of a flow. For instance, consider a QoE based packet marking scheme proposed by [122], where different video packets of a flow are assigned different priority classes based on packets' contribution to the user perceived video quality. For such packet marking schemes, standardized QCIs cannot be used because different performance characteristics are required for QoE based priority classes. The 3GPP TS 29.212 [128] has non-standard QCIs, 128-254, which are also known as operator specific QCIs. RAN and the LTE core network have been enhanced to support operator specific QCIs [128]. These QCIs allow network operators to define their own packet forwarding treatment, in terms of delay budget, admission thresholds, packet loss rate, etc., for QoE-aware partitioned priority classes.

TABLE II: Channel and QoS Parameters and Related Information Used by State-of-The-Art Content-Aware Downlink Scheduling Approaches.

Scheduling Approaches	**Channel Parameters**		**QoS Parameters**					**Extra Information Used**				
	CSI	*Average data rate*	*Target rate*	*Queue size*	*Traffic type*	*Target delay*	*HoL delay*	*Video complexity*	*Video coding type*	*Video rate adaptation*	*Analytical tools*	*Tech.*
Quality driven scheduling approaches												
[36], [37]	X				Video streaming			packet ordering of the encoded video according to their relative contribution to the final quality of the video	Non-scalable video (MPEG)	video packet dropping	Distortion model, Lagrangian multipliers, Convex optimization	HSDPA
[46]	X		X		Composite			I-based and P-based packets from the APP layer, and scheduling packets according to their importance	Non-scalable video (AVC)	video packet dropping	Linear precoding	OFDMA
[39–43], [60], [62]	X				Video streaming				Non-scalable video	video packet dropping		OFDMA
[44], [45], [47], [63]	X				Video streaming	X	X	each video frame is partitioned into one or more slices, each slice header acts as resynchronization marker to allow independent decoding to the slices, and each slice contains a number of I macro-blocks	Non-scalable video	video packet dropping	Gradient distortion projection, Game theory	OFDMA
[48]	X				Video streaming			source packets prioritizing in the queue based on their estimated impact on the overall video quality	Scalable video (H.264/SVC)	video packet dropping	Distortion model, Nakagami-m channel model, Lagrangian multipliers, Convex optimization	HSDPA
[49], [50]	X			X	Video streaming	X	X	utilization of the temporal and quality scalability of the video coding	Scalable video (H.264/SVC)	video packet dropping	Distortion function, Delay function, Lagrangian dual relaxation	OFDM
[51]	X				Video streaming	X	X	frames dependency on a video sequence and it is computed based on the video quality metrics, and the streams are organized in layers where upper layers cannot be decoded unless lower layers are correctly received	Scalable video (H.264/SVC)	video packet dropping	video frame significance, Cooperative game theory, Nash bargaining solution	OFDMA
[121]	X	X			Video streaming	X	X		Scalable video (H.264/SVC)	video packet dropping		OFDMA
[121]	X	X			Composite	X	X		Scalable video (H.264/SVC)	video packet dropping		OFDMA
[52]	X		X		Composite	X	X	dependency, temporal and quality IDs of the SVC video stream, and different compression ratios	3D Scalable video (H.264/SVC)	Video packet dropping	Simulation	OFDMA
Proxy driven radio resource allocation approaches												
[71], [73], [74]	X				Video streaming				Non-scalable video (Mpeg4 and AVC)	Adapting source rate at the server side	real-time testbed, and Gilbert-Elliot (GE) channel model	OFDM
[72], [75]	X				Composite				Non-scalable (AVC)	Adapting source rate at the server side	GE channel model	OFDMA
[83–86]	X				Composite				Non-scalable (AVC)	Transcoding		HSDPA
[87], [88]	X				Video streaming				SVC	SVC profile selection	Integer liner program, Multiple choice Knapsack problem, Fully polynomial time approximation scheme (FPTAS), Profit scaling technique, water filling	OFDMA
Client driven approaches												
QoE-DASH [93]	X	X			HTTP video			using different quality-rate for various video files	MPEG DASH		gradient based greedy algorithm	DASH-LTE
RAGA [94]	X	X	X	X	HTTP video				MPEG DASH		token-based mechanism to enforce the re-buffering constraints	DASH-LTE
Video-aware controller [95]	X	X	X	X	HTTP video				MPEG DASH		A central intelligence controller is placed at the core network for video-aware resource management	DASH-LTE

C. QCI modification

Under congested network, a video streaming bearer assigned with a low QoS profile, for instance QCI 9, may result in a poor user-perceived quality. In order to resolve this issue, the 3GPP has proposed and standardized bearer modification process. The bearer modification allows network operators to increase the priority, GBR or non-GBR bearer, of the UE with a higher QoS profile. In order to take advantage of the bearer modification, 3GPP TR36.933 has proposed an innovative solution to improve the QoE of the streaming users. According to the proposed solution, the bearer modification request can also be initiated by UEs, *i.e.*, UEs can request a higher QoS profile based on the video context. For instance if the playout buffer is depleting, then the UE can request a higher QoS profile by making a bearer modification request, so that increase in radio resource allocation lowers the video stalling probability.

D. DASH optimization

The major issue related with DASH streaming over LTE is the inaccurate throughput prediction for the next segment. Generally, DASH clients predict throughput of the next segment based on the achieved throughput of the previously downloaded segments. Such prediction deems to be inaccurate because the user may not have any knowledge about the underlying network conditions, as in a dynamic wireless environment, the network conditions can abruptly change. Herein, requesting a low data-rate representation can result in lower QoE whereas, requesting a high data-rate video content can result in constant interruption due to video stalling. In [126], the 3GPP proposed a solution to this problem which comprises a RAN assisted rate selection, *i.e.*, predicted throughput is provided by the eNodeB. According to the solution, the DASH client provides the available representations of the target segment to the eNodeB. The eNodeB computes the available downlink throughput for the DASH client based on congestion status, channel quality and available DASH representations. Finally, the predicted throughput is sent to the DASH client. With the predicted throughput, the client can select the rate of the next segment more accurately.

V. Simulation Set-up and Comparative Performance Analysis

This section analyzes, through simulations, the performance of important content-aware and content-unaware scheduling strategies. The simulation environment is also presented.

A. Simulation scenario

We consider a single cell scenario in which the serving eNodeB is at the center of the cell. The serving eNodeB's MAC scheduler controls all the available PRBs by allocating them to the active flows competing for resources. 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, and temporally organised in units which can be decoded independently from each other, each referred as GOP. The video sequences are encoded with the SVC codec (Medium Grain Scalability (MGS) [120]) and comprise a base layer and 12 quality layers. MGS scalability provides sufficient bit-rate granularity for rate adaptation. Each user is assigned a queue at the eNodeB. The set of packets in the buffer for each video streaming 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.

SVC video streaming flows have different priority packets, with the base layer packets contributing largest to the video quality. The increase in perceived video quality (MOS score) along with the addition of each quality enhancement layer is shown in Figure 8. In this work, MOS score is estimated by employing a full reference objective metric. Section V-B reports the adopted methodology for the computation of MOS score.

The packets entering the buffer are time stamped by the scheduler. The scheduler should assign enough resources to schedule the packets before the delay budget. Packets violating the delay budget are dropped from the queue; we assume that such packets have missed the decoding deadline at the receiving terminal.

LTE is a multicarrier system where radio resources are spread in time and frequency domains. Defining a scheduling strategy on a per-PRB basis, as shown in Figure 9, is simpler to implement. According to the figure, the user with the highest metric is allocated a PRB. The metric is either based on content-unaware strategies, such as QoS-aware rules, or video quality driven scheduling rules. In order to investigate the performance of the aforementioned approaches, an LTE link-level simulator built on MATLAB's object oriented features [130] is selected as the simulation platform. The wireless simulation parameters are reported in Table III. Our main goal is to analyze the performance of the selected strategies under different load scenarios. Specifically, we consider 5 different load scenarios with 8, 12, 16, 20 and 24 video streaming users. Initially, we simulate a network with 8 video streaming users with 2 *Ice*, 2 *News*, 2 *Soccer* and 2 *Crew* streaming sequences corresponding to an input average traffic rate of 7 Mbps. According to the simulation parameters, the average system capacity is approximately 7 Mbps (2.33 bits/sec/Hz considering a 3 MHz bandwidth and 8 video streaming users). The system load is then increased by adding 1 video streaming user from each of the considered video sequences until the total number of video streaming users in the network is 24. We simulate the following strategies:

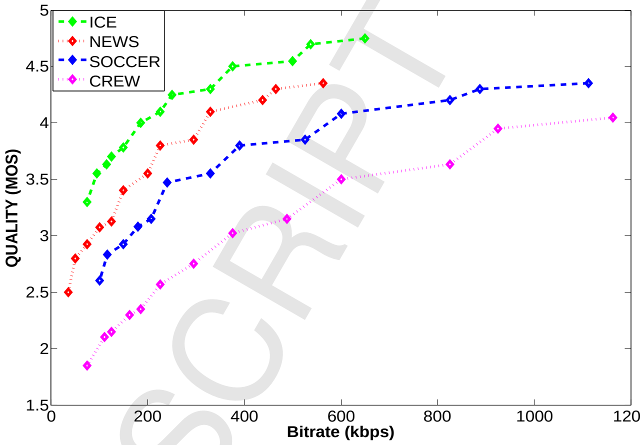


Fig. 8: Quality for the considered video sequences as SVC quality layers are added (bit-rate increased). MOS is derived by utilizing the VQM to MOS mapping [129].

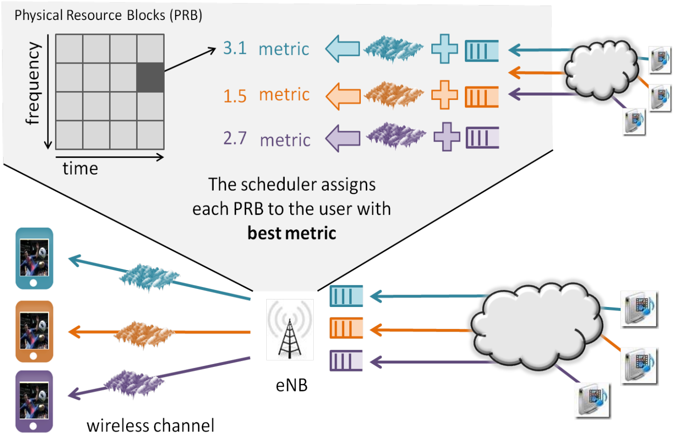


Fig. 9: Scheduling strategy on a per-PRB basis by computing a priority metric of each flow on a PRB [98].

- Strategy A (QoS unaware strategy): The PF scheduling rule is simulated by calculating the scheduling metric on each PRB as shown in Figure 9. This rule is simulated according to the guidelines given in [18].
- Strategy B (QoS aware strategy): The M-LWDF scheduling rule is simulated by computing the delay based scheduling metric on each PRB as done in Strategy A. A

TABLE III: 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)
Delay budget	500 TTIs (1 TTI = 1ms)
Video streaming duration	15 sec

detailed implementation of the M-LWDF rule is given in [18].

- Strategy C (Video quality driven and delay-blind scheduling): Simplified Fine Granularity (SFG) [39] scheduling based on packet's contribution towards video quality is simulated. This strategy is similar to the ones proposed in [40], [41], [43]. The algorithm comprises a two step process at each scheduling epoch. In the first step, the scheduler sorts the packets of a flow based on the ratio of each packet contribution towards video quality and its size. In the second step, 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 [39].
- Strategy D (Video quality driven and delay aware scheduling): Strategy D considers video packet's quality contribution, similar to strategy C, with the exception that strategy D is delay aware. In the first step, strategy D performs sorting based on packet's video quality contribution and delay requirements. In the second step, the scheduling metric is a function of video quality as well as the approaching deadline of the video packets as reported in [50].
- Strategy E (Proxy driven radio resource allocation): In strategy E, the proxy responds to the dynamic wireless channel and congestion by performing rate adaptation as reported in [85–87]. In order to perform rate adaptation, the proxy considers the rate-quality trade-off model of the video streaming sequences (as shown in Figure 8), 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. In the literature, the radio resource allocation cycle ranges from 10 ms [88] to 1 sec [85], [86]. We assume that proxy based resource allocation decisions are taken every 100 ms. The proxy receives periodic, every 100 ms, congestion and channel quality information from the MAC layer of the eNodeB. The eNodeB utilizes a QoS aware M-LWDF packet scheduler.

B. MOS estimation

In order to estimate the perceived video quality performance of diverse scheduling rules, we utilize an objective metric known as Video Quality Metric (VQM) [129]. VQM is a full reference metric which is dependent on several objective parameters for measuring the perceptual effects of a wide range of impairments in the spatial and temporal dimensions. The metric comprises a linear combination, *i.e.*, weighted sum, of seven independent video quality parameters. Four parameters are based on spatial gradients, two parameters are a function of features extracted from chrominance components, and one parameter is a measure of motion and contrast [129]. The VQM estimates the perceived video quality difference, ranging from 0 (no impairment) to 1 (severely distorted), between the original and the processed videos. In order to analyze the correlation performance of the VQM with the subjective MOS, the Video Quality Expert Group (VQEG) [131] tested the model considering 1536 subjectively rated video sequences. According to the test, the metric gave an outstanding performance and achieved a correlation performance of 0.938. The metric was selected in ITU recommendations [132] and was standardized by the American National Standards Institute (ANSI) [129]. In this work, we estimate the subjective video quality, MOS, in terms of VQM.

$$MOS = 5 - (4 \times VQM) \tag{1}$$

The MOS ranges from 1 (very bad) to 5 (excellent), *i.e.*, for no perceived impairments the VQM value of 0 corresponds to the highest MOS, whereas for severely distorted videos the VQM value of 1 achieves the lowest MOS. Figure 8 reports, for 4 different video sequences, the increase in the estimated MOS score as a consequence of the addition of each quality layer. The MOS score is estimated by the VQM mapping reported in (1). For instance the *crew* video sequence with the base layer and 12 quality enhancement layers results in a VQM of 0.25. According to the VQM to MOS mapping in (1), the estimated MOS is 4 as shown in Figure 8.

C. Performance metrics

In order to evaluate the video quality performance of the nominated scheduling strategies, we consider the following metrics:

- Sum throughput of all the video streaming flows:

$$\text{Sum Throughput} = \left(\sum_{i=1}^{I} \frac{1}{n_{\mathrm{f}_i} - n_{\mathrm{s}_i}} \sum_{m=n_{\mathrm{s}_i}}^{n_{\mathrm{f}_i}} P_{\mathrm{t}_i}^{(m)} \right) \tag{2}$$

  where n_{s_i} and n_{f_i} are the starting and finishing time intervals of the streaming of flow i respectively. $P_{\mathrm{t}_i}^{(m)}$ is the size of flow i's successfully scheduled/transmitted packet, and I is the total number of video streaming flows.
- System PLR, by considering the successfully scheduled and dropped packets. Mathematically, it is given as:

$$\text{System PLR} = \frac{\sum_{i=1}^{I} \sum_{m=n_{\mathrm{s}_i}}^{n_{\mathrm{f}_i}} P_{\mathrm{d}_i}^{(m)}}{\sum_{i=1}^{I} \sum_{m=n_{\mathrm{s}_i}}^{n_{\mathrm{f}_i}} \left(P_{\mathrm{t}_i}^{(m)} + P_{\mathrm{d}_i}^{(m)} \right)} \tag{3}$$

where $P_{\mathrm{d}_i}^{(m)}$ is the size of the dropped packet of flow i at the eNodeB.

- Average MOS across the flows. The aforementioned metrics are network centric, *i.e.*, they are QoS based network measurements. In order to measure the performance of the scheduling strategies in terms of QoE, we utilize MOS derived from the VQM as discussed in Section V-B. The contribution of each SVC layer in terms of MOS is shown in Figure 8. For instance, if a strategy schedules all the layers of the *CREW* sequence to a user, this results in a MOS score of 4 as shown in Figure 8. On the other hand, if the scheduler schedules the base layer and only 9 quality enhancement layers, then the MOS score for the user decreases to 3.5. The average QoE performance of a strategy is computed in terms of average MOS.

$$\text{Average MOS} = \frac{\sum_{i=1}^{I} MOS_i}{I} \qquad (4)$$

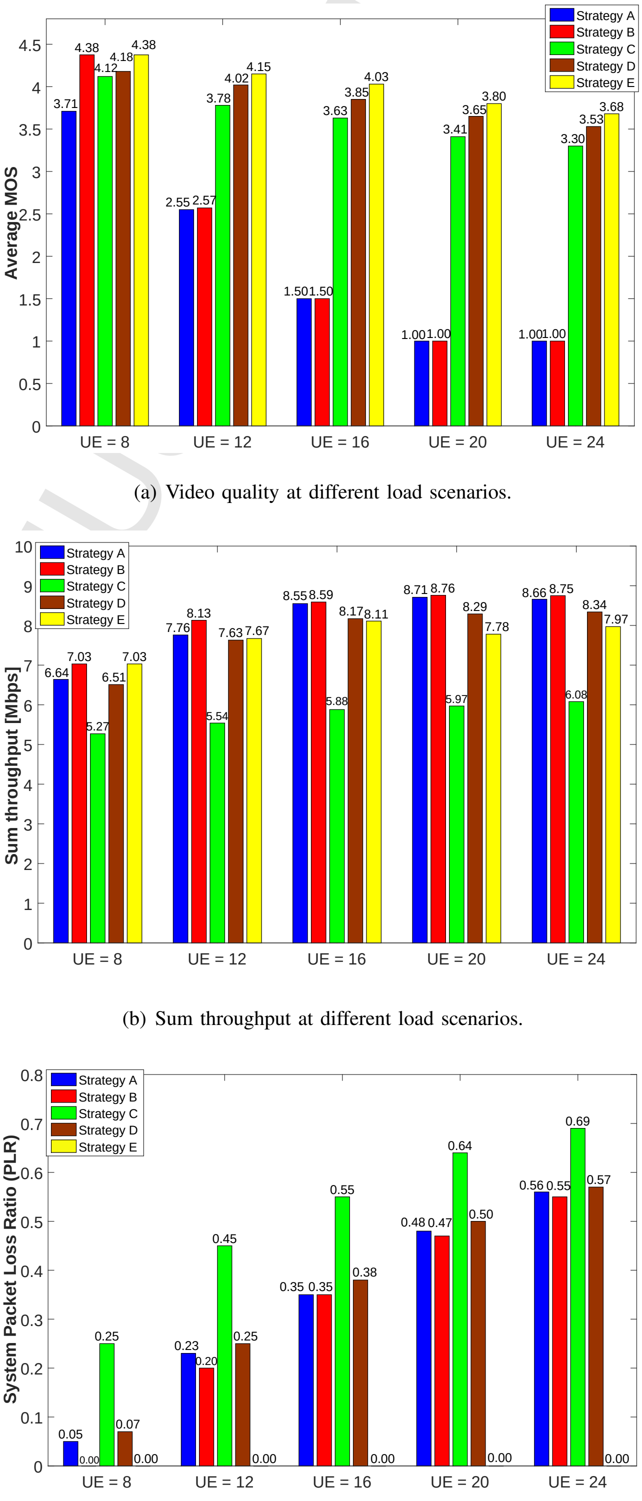
where MOS_i is the estimated MOS of flow i.

- System capacity in terms of total number of satisfied video streaming users, *i.e.*, video streaming users receiving a minimum average MOS of 3.

D. Results

The performance of the considered strategies in terms of average MOS, sum throughput and system PLR is shown in Figures 10(a), 10(b) and 10(c) respectively. Figure 11 shows the total number of satisfied users for each of the considered strategies under all the load scenarios.

According to Figures 10(b) and 10(c), both Strategy A (PF rule) and Strategy B (M-LWDF rule) perform best in terms of sum throughput and system PLR. However, the average MOS performance shows a rapid decline, as compared to other strategies, when the number of video streaming users is increased. It is important to note that each video streaming user has a buffer at the MAC layer of the eNodeB and the buffer is served according to the FIFO rule. The increase in the number of video streaming users increases the input traffic above the system capacity, which leads to an increase in the waiting time of the video packets buffered at the eNodeB. Under congestion, the less important video packets (quality enhancement layers) residing in the buffer till the delay bound block packets of the base layer (high priority packets). This phenomenon is also known as head of line blocking. When the head of line packets belong to the quality enhancement layers then, under congestion, base layer packets have to wait to be scheduled till the enhancement layer packets are either dropped (owing to delay bound violation) or scheduled from the buffer. This phenomenon increases the probability of delay bound violation of base layer packets. The probability of successful decoding of quality enhancement layers is significantly reduced, when the base layer video packets are dropped, resulting in a decrease in video streaming quality. When the system is left to run under high load by increasing the number of video streaming flows, the quality performance of the existing users in the network is violated resulting in an increase in the number of unsatisfied users as reported in



(a) Video quality at different load scenarios.

(b) Sum throughput at different load scenarios.

(c) System PLR at different load scenarios.

Fig. 10: Results comparison for different selected strategies.

Figure 11. Therefore, all the video quality blind scheduling rules must ensure that the arrival traffic should not exceed the wireless system capacity in terms of bits/sec. An admission control policy should block further video streaming flows from entering the system, under Strategy A and Strategy B, once the input traffic reaches the system capacity. In the considered scenario, the PF and the M-LWDF rules can accommodate 6 and 8 video streaming users respectively subject to the packet delay and wireless channel constraints. Hence both the PF and M-LWDF and other scheduling rules of this class require a strict flow based admission control policy, which inhibits the admission of further video streaming flows once the input traffic reaches the system capacity.

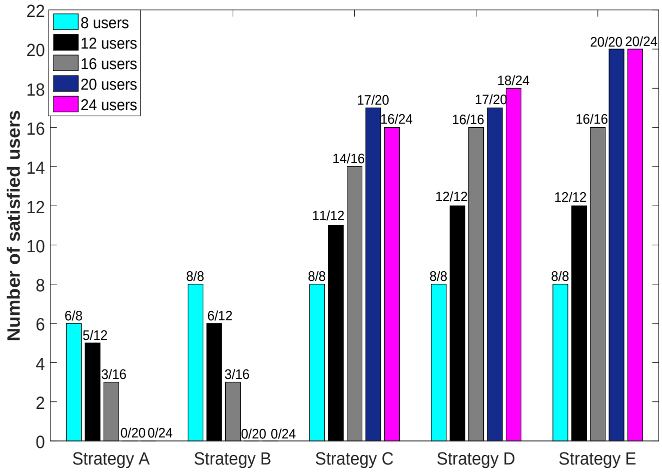


Fig. 11: Number of satisfied users under different load scenarios for different strategies.

Strategy C performs worst in terms of sum throughput and system PLR as shown in Figure 10(b) and 10(c). However, the average MOS performance increases substantially as compared to Strategy A and Strategy B as reported in Figure 10(a). It is important to note that Strategy C comprises a two step policy. In step 1, the most important video packet is selected for a flow. This step is also referred to as intra-flow scheduling. In the next step, each flow's scheduling priority is computed by considering the importance of the packet (the packet of a flow selected in step 1) in terms of video quality. This step is also referred to as inter-flow scheduling. The prioritization of important video packets in step 1 reduces their probability of delay bound violation. This step reduces the HoL blocking of important video packets which is not the case in Strategy A and Strategy B. Therefore, base layer video packets are never dropped from the buffer. This increases the number of satisfied users with the increase in the number video streaming users as shown in Figure 11. In the considered scenario, strategy C can accommodate 17 streaming users with satisfactory video quality as compared to 8 users for QoS-aware scheduling rule (strategy B). Strategy D shows a considerable increase in the sum throughput and average MOS performance as compared to strategy C. The significant performance difference between the two strategies is mainly due to the fact that the scheduling rule of strategy C is packet delay agnostic, *i.e.*, it is unable to determine the scheduling urgency of packets nearing the maximum tolerable delay bound. However both Strategy C and Strategy D suffer from high computation complexity. The maximum number of satisfied users with strategy D, under the considered load scenario, is 18 as shown in Figure 11.

Figure 10(a) also reports the average MOS performance of the proxy based resource allocation scheme (Strategy E) with M-LWDF scheduler at the eNodeB. Strategy E performs the best in terms of perceived video quality as shown in the figure. 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 system capacity, the proxy drops the video layers contributing lowest to the video quality for the flows having poor channel quality. This causes the average waiting time of the buffered packets below the delay bound, thus avoiding the delay bound violation at the eNodeB which results in a PLR of 0% as shown in Figure 10(c). This strategy requires a detailed signaling mechanism between the scheduler and proxy. Furthermore, the position of the proxy within the RAN is also an open issue.

It is important to note that content-unaware strategies require admission control which block flows from entering the system. The admission control of flows avoids HoL packet delay blocking of base layer video packets and ensures that packet delay bound of existing flows in the network is not violated. This reduces the exploitation of the most important phenomenon in wireless systems, *i.e.*, multi-user channel diversity. For instance, Figure 10(b) shows an increase in the sum throughput performance of Strategy A and B, but the system capacity in terms of satisfied users is decreased as reported in Figure 11. On the other hand, content-aware strategies do not require an admission control policy which leads to the exploitation of multi-user channel and content diversities. The increase in the system load leads to an increase in the sum MOS performance as shown in Figure 10(a) which enhances the system capacity in terms of the number of satisfied users as shown in Figure 11.

1) Complexity analysis of the considered strategies: In this section, we analyze the complexity of each of the considered strategies in terms of the maximum number of required iterations. Table IV summarizes the complexity and QoE performance of all the considered strategies. For strategy A and B, the scheduler computes $I \cdot M_{\text{PRB}}$ metrics per scheduling epoch, where I is the number of flows and M_{PRB} is the number of PRBs. The scheduling function for Strategy B requires the computation of HoL delay which comprises the recording of packets' arrival time at the eNodeB. However, this does not affect the number of iterations required to compute the scheduling rule. Therefore, the per-PRB scheduling rule has a linear dependency on the number of PRBs and flows for strategy A and B. Strategy C comprises a two step scheduling policy, where the first step requires sorting of packets and the second one consists of PRB assignment The step one of strategy C requires $n \cdot I \cdot \log(n \cdot I)$ iterations, where n is the number of packets in the buffer of a flow. Furthermore, step two requires $n \cdot I \cdot M_{\text{PRB}}$ iterations. Therefore, the total complexity per scheduling epoch is $O[n{\cdot}I{\cdot}\log(n \cdot I){+}n{\cdot}I{\cdot}M_{\text{PRB}}]$ which is a considerable increase as compared to $O(I \cdot M_{\text{PRB}})$ for strategy A and B. The sorting of packets in strategy D is based on each packet's quality contribution as well as delay requirements. Therefore, the strategy requires a computation of a priority function which is based on quality and delay. This step requires an additional $I \cdot n$ iterations which increases the complexity to $O[n \cdot I + n \cdot I \cdot \log(n \cdot I) + n \cdot I \cdot M_{\text{PRB}}]$.

The main goal of the proxy is to keep the input arrival traffic, at the MAC layer of the eNodeB, within the achievable rate region. The proxy based operation is performed at a uniform time interval, *i.e.*, 100 ms cycle. The proxy based strategy has three important steps. The first step estimates whether the input traffic is above the system capacity. This step computes the system capacity, available downlink throughput, and the system load, total enqueued packets, by considering the buffer status and average channel conditions of all the users. The maximum number of iterations required for this step is I. The second step is computed only if the system load is above the system capacity. This step comprises a resource allocation strategy, which is based on greedy algorithm. The main objective of the greedy algorithm is the maximization of the utility function, *i.e.*, sum MOS maximization. The algorithm is initiated by assigning an equal amount of resources for every user. At each subsequent iteration, a small amount of resources is taken from the user contributing least to the utility function and assigned to the user which maximizes the utility function. The process is repeated until there is no further improvement in the utility function. The worst case complexity of this step is $O(I^2)$. The final step of the proxy based strategy is the rate adaptation of each user. In this step, the radio resource allocation decisions are utilized to compute the downlink throughput of each user. The scalable video layers of each user are dropped according to the assigned throughput. This step requires $I \cdot N_l$ iterations, where N_l is the total number of scalable layers of each user.

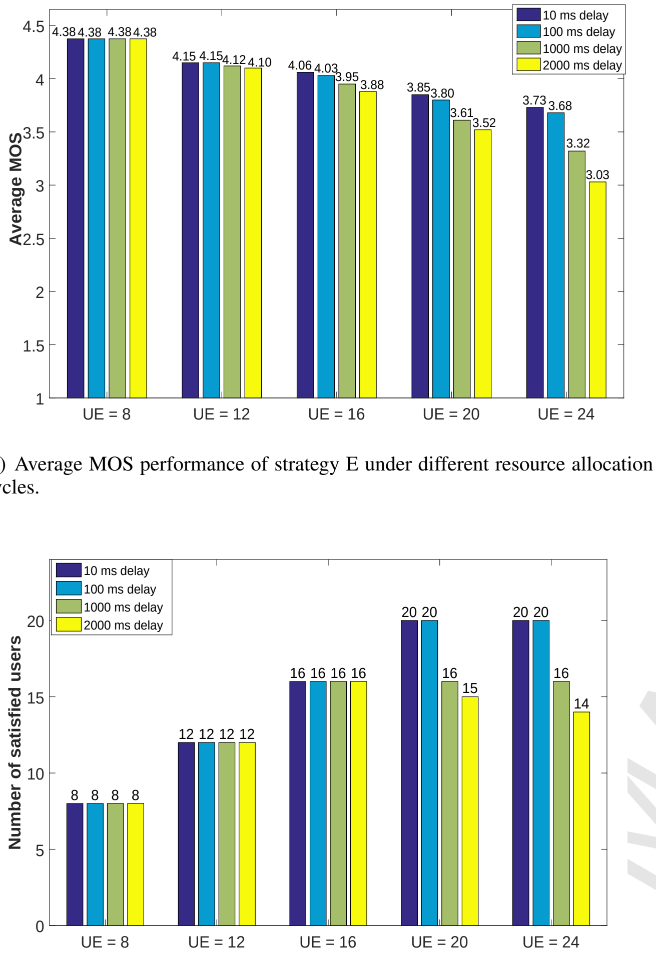
In order to study the impact of computation complexity on the QoE performance, we simulate strategy C, D and E with fewer iterations. For instance, strategy C requires a complexity of $n{\cdot}I{\cdot}\log(n \cdot I){+}n{\cdot}I{\cdot}M_{\text{PRB}}$ iterations. We simulate strategy C such that the sorting step requires only $n{\cdot}I{\cdot}\log(n)$ iterations, *i.e.*, the sorting is done on each queue independently without considering the quality contribution of the packets of other flows. Furthermore, the number of iterations of the second step is reduced to $I \cdot M_{\text{PRB}}$, *i.e.*, the scheduling metric is computed on each PRB by considering the quality contribution of the sorted packets of each flow as reported in [133] [134]. Similarly the scheduling complexity of strategy D is reduced to $O[n \cdot I + n \cdot I \cdot \log(n) + I \cdot M_{\text{PRB}}]$. In order to analyze the performance of strategy E with fewer iterations, we simulate the sum MOS maximization algorithm according to [135], which results in only $\frac{I+I^2}{2}$ iterations. According to Table IV, there is a considerable reduction in the QoE performance of strategy C and D. The reduction in the number of iterations significantly degrades the average QoE performance. However, the reduction in the complexity of strategy E has marginal impact on the average QoE performance as shown in Table IV. Therefore, the concept of proxy based strategy, where scheduling and resource allocation are performed separately results in low complexity and better QoE performance. On other hand, joint scheduling and resource allocation requires very high complexity in terms of the number of iterations as shown in Table IV.

TABLE IV: Complexity analysis of the considered stratgies.

Strategy	**Computation complexity at the scheduler**	**Other complexity**	**Average MOS (24 users)**
A	$O(I \cdot M_{\text{PRB}})$	none	1
B	$O(I \cdot M_{\text{PRB}})$	none	1
C	$O[n \cdot I \cdot \log(n \cdot I) + n \cdot I \cdot M_{\text{PRB}}]$	content information at eNodeB	3.30
D	$O[n \cdot I + n \cdot I \cdot \log(n \cdot I) + n \cdot I \cdot M_{\text{PRB}}]$	content information at eNodeB	3.53
E	$O(I \cdot M_{\text{PRB}})$	Proxy required with complexity $O(I + I^2 + I \cdot N_l)$	3.68
Reduced complexity			
C	$O[n \cdot I \cdot \log(I) + I \cdot M_{\text{PRB}}]$	content information at eNodeB	3.10
D	$O[n{\cdot}I{+}n{\cdot}I{\cdot}\log(I){+} I \cdot M_{\text{PRB}}]$	content information at eNodeB	3.25
E	$O(I \cdot M_{\text{PRB}})$	Proxy required with complexity $[O(\frac{I+I^2}{2}){+}I{\cdot}N_l]$	3.65

2) *Impact of delay on QoE performance for strategy E:* The aforementioned analysis shows that the proxy based strategy achieves the maximum number of satisfied users. Furthermore, it maximizes the QoE with low scheduling complexity. It is important to note that the proxy operates at a fixed time interval of 100 ms, *i.e.*, radio resource allocation of each user is computed after a delay of 100 ms. The performance of the proxy based strategy depends upon the changes in channel conditions between the time proxy receives the channel quality information and the time resource allocation decisions are computed for each flow. As a result, the decisions taken by the proxy may become obsolete because of the changes in the channel quality of each UE. In order to study the impact of delay on QoE, we simulate strategy E with four different radio resource allocation cycles. The simulation scenario is the same as reported in Section V-A with *Typical Urban* as the channel model. The average MOS performance and the total number of satisfied users, with resource allocation cycles of 10 ms, 100 ms, 1000 ms and 2000 ms, are shown in Figure 12. The four resource allocation cycles correspond to different positions of proxy. For instance, the delay of 10 ms refers to the position of proxy at the eNodeB, whereas delays of 100 ms and 1000 ms assume the position of proxy at the MME and P-GW respectively. Delays higher than 1s assume the location of proxy outside the LTE network.

According to the results, we observe that the impact of delay is more predominant with the increase in the total number of users. At lower system load, 8 and 12 users, the difference in the QoE performance is negligible. Furthermore, the difference in the average MOS performance for resource allocation cycles of 10 ms and 100 ms, 8 to 24 users, is minimal as shown in Figure 12(a). It is important to note that strategy E has two important entities, the scheduler and the



(a) Average MOS performance of strategy E under different resource allocation cycles.

(b) Number of satisfied users under strategy E with different resource allocation cycles.

Fig. 12: QoE performance of strategy E with resource allocation cycles of 10 ms, 100 ms, 1000 ms and 2000 ms.

proxy. The proxy avoids congesting the scheduler by dropping SVC layers of the UEs experiencing poor channel quality. Therefore, at lower system load, the impact of delay on QoE is negligible as the scheduler is not congested. The operation of the proxy becomes critical when the system becomes congested. According to Figure 12(a), there is a considerable decrease in the average MOS performance for the higher load scenarios with resource allocation cycles of 1000 ms and 2000 ms. The urban environment, *Typical Urban* channel model, has variations in the link quality due to the multipath fading, shadowing, and Doppler effects. The resource allocation decisions by the proxy becomes outdated due to the channel quality variations. According to Figure 12(b), the resource allocation cycles of 1000 ms and 2000 ms can accommodate only 16 and 15 satisfied users respectively. On the other hand, resource allocation cycles of 10 ms and 100 ms results in the same number of satisfied users. Therefore, QoE maximization for the proxy based strategies can be achieved by limiting the radio resource allocation cycle to 100 ms. This corresponds to the position of the proxy either close to the eNodeB, *i.e.*, MME node, or within the eNodeB. In such a scenario, the proxy and the scheduler can exchange information with little delay, which results in better QoE performance as shown in figure 12.

VI. Conclusion

This paper introduces a comprehensive literature review of the recent downlink scheduling approaches that mainly tackle the issues involved in video optimization. We categorize the downlink scheduling approaches into two broad classes: content-aware and content-unaware scheduling strategies. These classes are further divided into subclasses as shown in our proposed taxonomy. The subclasses are based on their technical contributions to the scheduling strategies available in the literature. For instance, QoS aware approaches are further categorized based on the type of QoS parameters, e.g., delay, packet loss rate, queue size, etc. The subclasses of content-aware strategies differ on how QoE based video optimization is performed in an LTE network. In the quality driven scheduling approach, for instance, the MAC layer scheduler performs objective video quality based optimization. On the other hand, proxy driven approach utilizes a content-unaware scheduler at the MAC layer, whereas content based radio resource allocation is performed at the proxy, which can either be located at the RAN or LTE core network. Furthermore, we analyze and compare different classes of scheduling in terms of QoS and QoE evaluation metrics. According to the simulation results, QoS-aware strategies maximize the system throughput but perform poorly in terms of user perceived video quality. On the other hand, the QoE-aware proxy based strategy maximizes the system capacity in terms of the total number of satisfied users and appears to be the most appealing strategy for an LTE downlink.

References

[1] “Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update 2016-2021,” White Paper, Cisco, February 2017.

[2] ITU-T Recommendation, E.800, “Definitions of terms related to quality of service,” International Telecommunication Union, Tech. Rep., 2008.

[3] ETSI TR 102 643 V1.0.1, “Human Factors (HF); Quality of Experience (QoE) requirements for real-time communication services,” European Telecommunications Standards Institute, Tech. Rep., 2009.

[4] ITU-R BT, Recommendation, “500-11, Methodology for the subjective assessment of the quality of television pictures,” International Telecommunication Union, Tech. Rep., 2002.

[5] Z.-N. Li, M. S. Drew, and J. Liu, *Fundamentals of multimedia*. Springer, 2004.

[6] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[7] K. Seshadrinathan, R. Soundararajan, A. Bovik, and L. Cormack, “Study of subjective and objective quality assessment of video,” *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1427–1441, June 2010.

[8] M. Razaak, M. Martini, and K. Savino, “A study on quality assessment for medical ultrasound video compressed via HEVC,” *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 5, pp. 1552–1559, September 2014.

[9] N. Barman and M. G. Martini, “H.264/MPEG-AVC, H.265/MPEGHEVC and VP9 codec comparison for live gaming video streaming,” in *Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, Erfurt, Germany, June. 2017.

[10] N. Barman, S. Zadtootaghaj, M. G. Martini, S. Moller, and S. Lee, "A comparative quality assessment study for gaming and non-gaming videos," in *Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, Serdinia, Italy, June. 2018.

[11] Y. Chen, K. Wu, and Q. Zhang, "From QoS to QoE: A Tutorial on Video Quality Assessment," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 2, pp. 1126–1165, 2015.

[12] M. M. Nasralla, C. Hewage, and M. Martini, "Subjective and objective evaluation and packet loss modeling for 3D video transmission over LTE networks," in *International Conference on Telecommunications and Multimedia (TEMU)*, Heraklion, Greece, July 2014, pp. 254–259.

[13] M. Martini, C. Hewage, M. Nasralla, R. Smith, I. Jourdan, and T. Rockall, "3-D robotic tele-surgery and training over next generation wireless networks," in *35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Osaka, Japan, July 2013, pp. 6244–6247.

[14] K. Zheng, X. Zhang, Q. Zheng, W. Xiang, and L. Hanzo, "Quality-of-experience assessment and its application to video services in LTE networks," *IEEE Wireless Communications*, vol. 22, no. 1, pp. 70–78, February 2015.

[15] M. G. Martini, C. W. Chen, Z. Chen, T. Dagiuklas, L. Sun, and X. Zhu, "Guest editorial QoE-aware wireless multimedia systems," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 7, pp. 1153–1156, 2012.

[16] T. Zhao, Q. Liu, and C. W. Chen, "QoE in video transmission: A user experience-driven strategy," *IEEE Communications Surveys & Tutorials*, 2016.

[17] P. V. Pahalawatta and A. K. Katsaggelos, "Review of content-aware resource allocation schemes for video streaming over wireless networks," *Wireless Communications and Mobile Computing*, vol. 7, no. 2, pp. 131–142, 2007.

[18] F. Capozzi, G. Piro, L. Grieco, G. Boggia, and P. Camarda, "Downlink packet scheduling in LTE cellular networks: Key design issues and a survey," *IEEE Communications Surveys and Tutorials*, vol. 15, no. 2, pp. 1–23, 2012.

[19] L. Skorin-Kapov, M. Varela, T. Hoßfeld, and K.-T. Chen, "A survey of emerging concepts and challenges for qoe management of multimedia services," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 14, no. 2s, p. 29, 2018.

[20] T. O. Olwal, K. Djouani, and A. M. Kurien, "A survey of resource management toward 5G radio access networks," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 3, pp. 1656–1686, 2016.

[21] S. Sesia, I. Toufic, and M. Baker, "LTE - the UMTS long term evolution," *Wiley, UK*, 2009.

[22] W. Tong, E. Sich, Z. Peiying, and J. Costa, "True broadband multimedia experience," *IEEE Micro-wave Mag.*, vol. 9, no. 4, pp. 64–73, Aug. 2008.

[23] S. Shakkottai, T. S. Rappaport, and P. C. Karlsson, "Cross-layer design for wireless networks," *IEEE Commun. Mag.*, vol. 41, no. 10, pp. 74–80, 2003.

[24] M. M. Nasralla, O. Ognenoski, and M. G. Martini, "Bandwidth scalability and efficient 2D and 3D video transmission over LTE networks," in *IEEE International Conference on Communications Workshops (ICC)*, Budapest, Hungary, June 2013, pp. 617 – 621.

[25] 3GPP Technical Specification 36.213, "Physical layer procedures (release 8)," www.3gpp.org.

[26] G. Piro, L. A. Grieco, G. Boggia, F. Capozzi, and P. Camarda, "Simulating LTE cellular systems: an open source framework," in *IEEE Trans. Veh. Technol.*, Los Angeles, USA, October, 2010, pp. 1–16.

[27] H. Ramli, R. Basukala, K. Sandrasegaran, and R. Patachaianand, "Performance of well known packet scheduling algorithms in the downlink 3GPP LTE system," in *Proceeding of IEEE Malaysia International Conference on Communications (MICC)*, Kuala Lumpur, Malaysia, December 2009, pp. 815–820.

[28] H. Jiang, W. Zhuang, and X. Shen, "Cross-layer design for resource allocation in 3G wireless networks and beyond," *IEEE Communications Magazine*, vol. 43, no. 12, pp. 120–126, Dec. 2005.

[29] R. A. Berry, , and E. Yeh, "Cross-layer wireless resource allocation," *IEEE Signal Processing Magazine*, vol. 21, no. 5, pp. 59–68, Sept. 2004.

[30] L. Georgiadis, M. J. Neely, and L. Tassiulas, "Resource allocation and cross-layer control in wireless networks," *Foundations and Trends in Networking*, vol. 1, no. 1, pp. 1401–1415, 2006.

[31] F. Fu and M. van der Schaar, "Decomposition Principles and Online Learning in Cross-Layer Optimization for Delay-Sensitive Applications," *IEEE Transactions on Signal Processing*, vol. 58, no. 3, pp. 1401–1415, Mar. 2010.

[32] T. Fingscheidt, T. Hindelang, R. V. Cox, and N. Seshadri, "Joint source-channel (de-)coding for mobile communications," *IEEE Transactions on Communications*, vol. 50, no. 2, pp. 200–212, 2002.

[33] T. Breddermann, H. Luders, P. Vary, I. Aktas, and F. Schmidt, "Iterative source-channel decoding with cross-layer support for wireless VoIP," in *IEEE International ITG Conference on Source and Channel Coding*, Siegen, Germany, Jan 2010, pp. 1–6.

[34] G. Liebl, T. Stockhammer, C. Buchner, and A. Klein, "Radio link buffer management and scheduling for video streaming over wireless shared channels," in *Proc. of the Packet Video Workshop*, Irvine, CA, USA, December 2004.

[35] G. Liebl, H. Jenkac, T. Stockhammer, and C. Buchner, "Radio link buffer management and scheduling for wireless video streaming," *Telecommunication Systems, Springer*, vol. 30, no. 1-3, pp. 255–277, November 2005.

[36] P. V. Pahalawatta, R. Berry, T. N. Pappas, and A. K. Katsaggelos, "A content-aware scheduling scheme for video streaming to multiple users over wireless networks," in *Proc. of IEEE European Signal Processing Conference*, Florence, September 2006, pp. 1–5.

[37] P. Pahalawatta, R. Berry, T. Pappas, and A. Katsaggelos, "Content-aware resource allocation and packet scheduling for video transmission over wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 25, no. 4, pp. 749–759, 2007.

[38] F. Li, G. Liu, J. Xu, and L. He, "Packet scheduling and resource allocation for video transmission over downlink OFDMA networks," in *Fourth International Conference on IEEE Communications and Networking in China (ChinaCOM)*, Xian, China, August 2009, pp. 1–5.

[39] Y. Zhang and G. Liu, "Fine granularity resource allocation algorithm for video transmission in orthogonal frequency division multiple access system," *IEEE IET (Institution of Engineering and Technology) Communications*, vol. 7, no. 13, pp. 1383–1393, September 2013.

[40] F. Li, P. Ren, and Q. Du, "Joint Packet Scheduling and Subcarrier Assignment for Video Communications Over Downlink OFDMA Systems," *IEEE Transactions on Vehicular Technology*, vol. 61, no. 6, pp. 2753–2767, July. 2012.

[41] F. Li, D. Zhang, and M. Wang, "Multiuser multimedia communication over orthogonal frequency-division multiple access downlink systems," *Concurrency and Computation: Practice and Experience*, vol. 25, no. 9, pp. 1081–1090, June 2013.

[42] F. Li, G. Liu, and L. He, "A Low Complexity Algorithm of Packet Scheduling and Resource Allocation for Wireless VoD Systems," *IEEE Transactions on Consumer Electronics*, vol. 56, no. 2, pp. 1057–1062, May 2010.

[43] P. Li, Y. Chang, N. Feng, and F. Yang, "A Cross-Layer Algorithm of Packet Scheduling and Resource Allocation for Multi-User Wireless Video Transmission," *IEEE Transactions on Consumer Electronics*, vol. 57, no. 3, pp. 1128–1134, Sept. 2011.

[44] F. Li, G. Liu, and L. He, "Application-driven cross-layer approaches to video transmission over downlink OFDMA networks," in *IEEE GLOBECOM Workshops*, Honolulu, November 2009, pp. 1–6.

[45] P. E. Omiyi and M. G. Martini, "Cross-layer content/channel aware multi-user scheduling for downlink wireless video streaming," in *5th IEEE International Symposium on Wireless Pervasive Computing (ISWPC)*, Modena, Italy, May 2010, pp. 412–417.

[46] S. Karachontzitis, T. Dagiuklas, and L. Dounis, "Novel cross-layer scheme for video transmission over LTE-based wireless systems," in *IEEE International Conference on Multimedia and Expo (ICME)*, Barcelona, Spain, July 2011, pp. 1–6.

[47] Z. Lu, X. Wen, W. Zheng, Y. Ju, and D. Ling, "Gradient projection based QoS driven cross-layer scheduling for video applications," in *IEEE International Conference on Multimedia and Expo (ICME)*, Barcelona, Spain, July 2011, pp. 1–6.

[48] E. Maani, P. V. Pahalawatta, R. Berry, and A. K. Katsaggelos, "Content-aware packet scheduling for multiuser scalable video delivery over wireless networks," in *SPIE Optical Engineering and Applications*, San Diego, CA, August 2009.

[49] X. Ji, J. Huang, M. Chiang, and F. Catthoor, "Downlink OFDM scheduling and resource allocation for delay constraint SVC streaming," in *IEEE International Conference on Communications (ICC)*, Beijing, China, May 2008, pp. 2512–2518.

[50] X. Ji, J. Huang, M. Chiang, G. Lafruit, and F. Catthoor, "Scheduling and resource allocation for SVC streaming over OFDM downlink systems," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, no. 10, pp. 1549–1555, 2009.

[51] N. Khan, M. G. Martini, and Z. Bharucha, "Quality-aware fair downlink scheduling for scalable video transmission over LTE systems," in *IEEE*

International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), Cesme, Turkey, June 2012, pp. 334–338.

[52] H. D. Appuhami, M. G. Martini, and C. T. Hewage, “Channel and content aware 3D video scheduling with prioritized queuing,” in *Wireless Advanced (WiAd)*, London, UK, June 2012, pp. 159–163.

[53] H. H. Juan, H. C. Huang, C. Huang, and T. Chiang, “Scalable video streaming over mobile WiMAX,” in *IEEE International Symposium on Circuits and Systems*, New Orleans, May 2007, pp. 3463–3466.

[54] P. Amon, T. Rathgen, and D. Singer, “File format for scalable video coding,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 9, pp. 1174–1185, September 2007.

[55] S. Wenger, Y. Wang, and T. Schierl, “Transport and signaling of SVC in IP networks,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 9, pp. 1164–1173, September 2007.

[56] B. Fu, D. Staehle, G. Kunzmann, E. Steinbach, and W. Kellerer, “QoE-aware priority marking and traffic management for H.264/SVC-based mobile video delivery,” in *8th ACM workshop on Performance monitoring and measurement of heterogeneous wireless and wired networks*, New York, USA, November 2013, pp. 173–180.

[57] M. M. Nasralla, “Video quality and QoS-driven downlink scheduling for 2D and 3D video over LTE networks,” Ph.D. dissertation, Kingston University, London, United Kingdom, December, 2015.

[58] N. Khan and M. G. Martini, “QoE-based video delivery over lte hierarchical architecture,” in *27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Valencia, Spain, Sept 2016, pp. 1–6.

[59] N. Khan, “Quality-driven multi-user resource allocation and scheduling over LTE for delay sensitive multimedia applications,” in *Ph.D. Dissertation*. Kingston University London, UK, 2014.

[60] R. Perera, A. Fernando, T. Mallikarachchi, H. K. Arachchi, and M. Pourazad, “QoE aware resource allocation for video communications over LTE based mobile networks,” in *10th International Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness*, Rhodes, Greece, August 2014, pp. 63–69.

[61] L. He, G. Liu, and C. Yuchen, “Buffer status and content aware scheduling scheme for cloud gaming based on video streaming,” in *IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, Chengdu, China, July 2014, pp. 1–6.

[62] F. Li, D. Zhang, and L. Wang, “Packet importance based scheduling strategy for H.264 video transmission in wireless networks.” Springer, 2014, pp. 1–17.

[63] Y. Ju, Z. Lu, D. Ling, X. Wen, W. Zheng, and W. Ma, “QoE-based cross-layer design for video applications over LTE.” Springer, 2014, vol. 72, no. 2, pp. 1093–1113.

[64] M. Rugelj, U. Sedlar, M. Volk, J. Sterle, M. Hajdinjak, and A. Kos, “Novel cross-layer QoE-aware radio resource allocation algorithms in multiuser OFDMA systems,” *IEEE Transactions on Communications*, vol. 62, no. 9, pp. 3196–3208, Sept 2014.

[65] M. Li, Z. Chen, P. H. Tan, S. Sun, and Y.-P. Tan, “QoE-aware video streaming for SVC over multiuser MIMO-OFDM systems,” *Journal of Visual Communication and Image Representation*, vol. 26, pp. 24–36, 2015.

[66] T. Ghalut, H. Larijani, and A. Shahrabi, “QoE-aware optimization of video stream downlink scheduling over LTE networks using RNNs and genetic algorithm,” *Procedia Computer Science*, vol. 94, pp. 232–239, 2016.

[67] V. F. Monteiro, D. A. Sousa, T. F. Maciel, F. R. M. Lima, and F. R. P. Cavalcanti, “A QoE-aware scheduler for OFDMA networks,” *Journal of Communication and Information Systems*, vol. 31, no. 1, 2016.

[68] A. A. Khalek, C. Caramanis, and R. W. Heath, “Delay-constrained video transmission: Quality-driven resource allocation and scheduling,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 9, no. 1, pp. 60–75, 2015.

[69] S. E. Ghoreishi and A. H. Aghvami, “Power-efficient QoE-aware video adaptation and resource allocation for delay-constrained streaming over downlink OFDMA,” *IEEE Communications Letters*, vol. 20, no. 3, pp. 574–577, 2016.

[70] M. M. Nasralla, M. Razaak, I. U. Rahman, and M. G. Martini, “Content-aware packet scheduling strategy for medical ultrasound videos over LTE wireless networks,” *Computer Networks*, vol. 140, p. 126–137, 2018.

[71] W. Kellerer, L. U. Choi, and E. Steinbach, “Cross-layer adaptation for optimized B3G service provisioning,” in *International Symposium on Wireless Personal Multimedia Communications (WPMC)*, Yokosuka, Japan, Oct 2003.

[72] S. Khan, Y. Peng, E. Steinbach, M. Sgroi, and W. Kellerer, “Application-driven cross-layer optimization for video streaming over wireless networks,” *IEEE Communications Magazine*, vol. 44, no. 1, pp. 122–130, 2006.

[73] Y. Peng, S. Khan, E. Steinbach, M. Sgroi, and W. Kellerer, “Adaptive resource allocation and frame scheduling for wireless multi-user video streaming,” in *IEEE International Conference on Image Processing (ICIP)*, September 2005.

[74] S. Shakkottai, T. S. Rappaport, and P. C. Karlsson, “Cross-layer design for wireless networks,” *IEEE Communications Magazine*, vol. 41, no. 10, pp. 74–80, 2003.

[75] S. Khan, S. Duhovnikov, E. Steinbach, and W. Kellerer, “MOS-based multiuser multiapplication cross-layer optimization for mobile multimedia communication.” Hindawi, 2007, p. 11.

[76] A. Saul, “Wireless resource allocation with perceived quality fairness,” in *IEEE Annual Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, USA, November 2008.

[77] B. J. Kim, “A network service providing wireless channel information for adaptive mobile applications. I. Proposal,” in *IEEE International Conference on Communications (ICC)*, Helsinki, Finland, June 2001, pp. 1345–1351.

[78] L. A. Larzon, U. Bodin, and O. Schelen, “Hints and notifications [for wireless links],” in *IEEE Wireless Communications and Networking Conference (WCNC)*, Orlando, Florida, USA, March 2002.

[79] A. Takacs, A. Kovacs, I. Godor, F. Kalleitner, H. Brand, M. Ek, T. Stefansson, and F. Sjoberg, “Journal of computers,” *The Layer-Independent Descriptor Concept*, vol. 1, no. 2, pp. 23–32, 2006.

[80] M. G. Martini, M. Mazzotti, C. Lamy-Bergot, J. Huusko, and P. Amon, “Content adaptive network aware joint optimization of wireless video transmission,” *IEEE Communications Magazine*, vol. 45, no. 3, pp. 1–10, 2007.

[81] J. Huusko, J. Vehkaperä, P. Amon, C. Lamy-Bergot, G. Panza, J. Peltola, and M. G. Martini, “Cross-layer architecture for scalable video transmission in wireless network,” *Signal Processing: Image Communication*, vol. 22, no. 3, pp. 317–330, 2007.

[82] M. G. Martini and V. Tralli, “Video quality based adaptive wireless video streaming to multiple users,” in *Proc. IEEE International Symposium on Broadband Multimedia Systems and Broadcasting*, Las Vegas, NV, April 2 2008, pp. 1–4.

[83] S. Thakolsri, S. Cokbulan, D. Jurca, Z. Despotovic, and W. Kellerer, “QoE-driven cross-layer optimization in wireless networks addressing system efficiency and utility fairness,” in *IEEE Workshop on Multimedia Communications and Services (GLOBECOM Workshops)*, Houston, USA, December 2011.

[84] S. Thakolsri, W. Kellerer, S. Khan, and E. Steinbach, “Application-driven cross layer optimization in wireless networks,” in *Seminar on Service Quality Evaluation in Wireless Netowrks supported by COST 290*, Stuttgart, Germany, June 2007.

[85] S. Thakolsri, S. Khan, E. Steinbach, and W. Kellerer, “QoE-driven cross-layer optimization for high speed downlink packet access.” *Journal of Communications*, vol. 4, no. 9, pp. 669–680, 2009.

[86] S. Thakolsri, W. Kellerer, and E. Steinbach, “Application-driven cross layer optimization for wireless networks using MOS-based utility functions,” in *International Conference on Communications and Networking in China (ChinaCOM)*, Xi An, China, August 2009.

[87] S. Cicalò and V. Tralli, “Distortion-Fair Cross-Layer Resource Allocation for Scalable Video Transmission in OFDMA Wireless Networks,” *IEEE Transactions on Multimedia*, vol. 16, no. 3, pp. 848–863, April. 2014.

[88] A. Ahmedin, K. Pandit, D. Ghosal, and A. Ghosh, “Exploiting scalable video coding for content aware downlink video delivery over LTE,” in *Distributed Computing and Networking*. Springer, 2014, pp. 423–437.

[89] T. Zhen, Y. Xu, T. Yang, and B. Hu, “QoE-aware proactive caching of scalable videos over small cell networks,” *arXiv preprint arXiv:1604.07572*, 2016.

[90] S. Akhshabi, A. C. Begen, and C. Dovrolis, “An experimental evaluation of rate-adaptation algorithms in adaptive streaming over HTTP,” in *Proceedings of the second annual ACM (Association for Computing Machinery) conference on Multimedia systems*, New York, USA, 2011, pp. 157–168.

[91] G. A. Tian and L. Yong, “On adaptive HTTP streaming to mobile devices,” in *20th International Packet Video Workshop (PV)*, San Jose, CA, December 2013, pp. 1–8.

[92] T. Thang, Q. Ho, J. Kang, and A. Pham, “Adaptive streaming of audio-visual content using MPEG DASH,” *IEEE Transactions on Consumer Electronics*, vol. 17, no. 9, pp. 78–85, 2012.

[93] E. A. Essaili, D. Schroeder, M. Shehada, W. Kellerer, and E. Steinbach, “Quality-of-experience driven adaptive HTTP media delivery,” in *IEEE*

International Conference on Communications (ICC), Budapest, Hungary, June 2013, pp. 2480–2485.

[94] V. Ramamurthi and O. Oyman, “Video-QoE aware radio resource allocation for HTTP adaptive streaming,” in *IEEE International Conference on Communications (ICC)*, Sydney, June 2014, pp. 1076–1081.

[95] V. Ramamurthi, O. Oyman, and J. Foerster, “Video-QoE aware resource management at network core,” in *IEEE Global Communications Conference (GLOBECOM)*, Austin, TX, December 2014, pp. 1418–1423.

[96] O. Ognenoski, M. M. Nasralla, M. Razaak, M. Martini, and P. Amon, “DASH-based video transmission over LTE networks,” in *IEEE International Conference on Communications (ICC) - Workshop on Quality of Experience-based Management for Future Internet Applications and Services (QoE-FI)*, London, UK, June 2015.

[97] M. Zhao, X. Gong, J. Liang, W. Wang, X. Que, and S. Cheng, “QoE-driven cross-layer optimization for wireless dynamic adaptive streaming of scalable videos over HTTP,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 3, pp. 451–465, 2015.

[98] N. Khan, M. M. Nasralla, and M. Martini, “Network and user centric performance analysis of scheduling strategies for video streaming over LTE,” in *IEEE International Conference on Communications (ICC) - Workshop on Quality of Experience-based Management for Future Internet Applications and Services (QoE-FI)*, London, UK, June 2015.

[99] R. Basukala, H. M. Ramli, and K. Sandrasegaran, “Performance analysis of EXP/PF and M-LWDF in downlink 3GPP LTE system,” in *Proceedings of IEEE International Conference on First Asian Himalayas*, Kathmandu, Nepal, November, 2009, pp. 1–5.

[100] M. G. Martini, C. T. Hewage, and M. M. Nasralla, “3D robotic surgery and training at a distance,” in *3D Future Internet Media*. Springer, 2014, pp. 257–274.

[101] S. J. Bae, B.-G. Choi, and M. Y. Chung, “Delay-aware packet scheduling algorithm for multiple traffic classes in 3GPP LTE system,” in *IEEE 17th Asia-Pacific Conference on Communications (APCC)*, Sabah,Malaysia, October 2011, pp. 33–37.

[102] B. Sadiq, R. Madan, and A. Sampath, “Downlink scheduling for multiclass traffic in LTE,” *EURASIP Journal on Wireless Communications and Networking*, p. 14, 2009.

[103] W. Shih-Jung, “A channel quality-aware scheduling and resource allocation strategy for downlink LTE systems,” *Journal of Computational Information Systems*, vol. 8, no. 2, pp. 695–707, 2012.

[104] M. Iturralde, A. Wei, T. Yahiya, and A. Beylot, “ Performance study of multimedia services using virtual token mechanism for resource allocation in LTE networks.” in *IEEE Vehicular Technology Conference (VTC)*, San Francisco, CA, September 2011, pp. 1–5.

[105] G. Monghal, K. I. Pedersen, I. Z. Kovacs, and P. E. Mogensen, “QoS oriented time and frequency domain packet schedulers for the UTRAN long term evolution,” in *IEEE Vehicular Technology Conference, (VTC)*, Singapore, May 2008, pp. 2532–2536.

[106] Y. Zaki, T. Weerawardane, C. Gorg, and A. Timm-Giel, “Multi-QoS-aware fair scheduling for LTE,” in *IEEE 73rd Vehicular technology conference (VTC)*, Yokohama, Japan, May 2011, pp. 1–5.

[107] D. N. Skoutas and A. N. Rouskas, “Scheduling with QoS provisioning in mobile broadband wireless systems,” in *IEEE European Wireless Conference (EW)*, Lucca, Italy, April 2010, pp. 422–428.

[108] M. M. Nasralla and M. G. Martini, “A downlink scheduling approach for balancing QoS in LTE wireless networks,” in *IEEE 24th International Symposium on PIMRC*, London, UK, September 2013, pp. 1571–1575.

[109] P. Svedman, S. K. Wilson, and B. Ottersten, “A QoS-aware proportional fair scheduler for opportunistic OFDM,” in *IEEE 60th Vehicular Technology Conference*, September 2004, pp. 558–562.

[110] K. Sun, Y. Wang, T. Wang, Z. Chen, and G. Hu, “Joint channel-aware and queue-aware scheduling algorithm for multi-user MIMO-OFDMA systems with downlink beamforming,” in *IEEE 68th Vehicular Technology Conference*, Calgary, September 2008, pp. 1–5.

[111] L. Yanhui, W. Chunming, Y. Changchuan, and Y. Guangxin, “Downlink scheduling and radio resource allocation in adaptive OFDMA wireless communication systems for user-individual QoS,” in *Proceedings of the World Academy of Science, Engineering and Technology*, 2006, pp. 221–225.

[112] M. Sarkar and H. Sachdeva, “A QoS aware packet scheduling scheme for WiMAX,” in *Proceedings of the World Congress on Engineering and Computer Science*, San Francisco, USA, October 2010, pp. 857–864.

[113] G. Piro, L. A. Grieco, G. Boggia, R. Fortuna, and P. Camarda, “Two-level downlink scheduling for real-time multimedia services in LTE networks,” *IEEE Transactions on Multimedia*, vol. 13, no. 5, pp. 1052–1065, May 2011.

[114] C. He and R. D. Gitlin, “Application-specific and QoS-aware scheduling for wireless systems,” in *IEEE 25th International Symposium on Personal, Indoor and Mobile Radio Communications*, Washington D.C., USA, September 2014.

[115] I. S. Comsa, R. Trestian, and G. Ghinea, “360^0 Mulsemedia experience over next generation wireless networks - A reinforcement learning approach,” in *International Conference on Quality of Multimedia Experience (QoMEX)*, Sardinia, Italy, May 2018, pp. 1–6.

[116] J. G. Choi and S. Bahk, “Cell-throughput analysis of the proportional fair scheduler in the single-cell environment,” *IEEE Transactions on Vehicular Technology*, vol. 56, no. 2, pp. 766–778, March 2007.

[117] H. A. M. Ramli, K. Sandrasegaran, R. Basukala, R. Patachaianand, and T. S. Afrin, “Video streaming performance under well-known packet scheduling algorithms,” *International Journal of Wireless & Mobile Networks (IJWMN)*, vol. 3, no. 1, pp. 25–38, February 2011.

[118] P. Ameigeiras, J. Wigard, and P. Mogensen, “Performance of the M-LWDF scheduling algorithm for streaming services in HSDPA,” in *IEEE conference on Vehicular Technology*, Los Angeles, USA, September 2004, pp. 999–1003.

[119] ITU-T Recommendation, P.910, “Subjective video quality assessment methods for multimedia applications,” International Telecommunication Union, Tech. Rep., 2008.

[120] H. Schwarz, D. Marpe, and T. Wiegand, “Overview of the scalable video coding extension of the H.264/AVC standard,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 9, pp. 1103–1120, September 2007.

[121] N. Khan, M. G. Martini, and D. Staehle, “Opportunistic Proportional Fair Downlink Scheduling for Scalable Video Transmission over LTE Systems,” in *IEEE Vehicular Technology Conference (VTC)*, Las Vegas, USA, Sept. 2013.

[122] B. Fu, D. staehle, G. Kunzmann, E. Steinbach, and W. Kellerer, “QoE-based SVC layer dropping in LTE networks using content-aware layer priorities,” *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 12, no. 1, p. 23, 2015.

[123] N. Khan and M. Martini, “Hysteresis based rate adaptation for scalable video traffic over an LTE downlink,” in *IEEE International Conference on Communications (ICC) - Workshop on Smart Communication Protocols and Algorithms*, London, UK, June 2015.

[124] S. Colonnese, F. Cuomo, T. Melodia, and I. Rubin, “A cross-layer bandwidth allocation scheme for HTTP-based video streaming in LTE cellular networks,” *IEEE Communications Letters*, vol. 21, no. 2, p. 386–389, 2017.

[125] *3GPP TS 36.300: Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description*, 3GPP Std. TS 36.300, 2016.

[126] *3GPP TR 36.933: Technical Specification Group Radio Access Network; Study on Context Aware Service Delivery in RAN for LTE* , 3GPP Std. TR 36.933, 2017.

[127] *3GPP TS 23.203: Policy Control and Charging architecture*, 3GPP Std. TS 23.203, 2018.

[128] *3GPP TS 29.212: Policy and Charging Control (PCC); Reference points*, 3GPP Std. TS 29.212, 2018.

[129] M. H. Pinson and S. Wolf, “A new standardized method for objectively measuring video quality,” *IEEE Transactions on Broadcasting*, vol. 50, no. 3, pp. 312–322, 2004.

[130] C. Mehlfuhrer, J. C. Ikuno, M. Simko, S. Schwarz, M. Wrulich, and M. Rupp, “The Vienna LTE simulators - Enabling reproducibility in wireless communications research,” *EURASIP Journal on Advances in Signal Processing*, vol. 2011, no. 29, pp. 1–14, July.

[131] VQEG FR-TV Phase II test, “Final report from the video quality experts group on the validation of objective models of video quality assessment,” Video Quality Expert Group, Tech. Rep., 2003.

[132] ITU-T J.144, “Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference,” International Telecommunication Union, Tech. Rep., 2004.

[133] N. Khan and M. G. Martini, “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*, vol. 2016, no. 93, pp. 1–23, 2016.

[134] N. Khan and M. Martini, “Hysteresis based rate adaptation for scalable video traffic over an LTE downlink,” in *IEEE International Conference on Communications (ICC) - Workshop on Smart Communication Protocols and Algorithms*, London, UK, June 2015.

[135] A. Saul and G. Auer, “Multiuser resource allocation maximizing the perceived quality,” *EURASIP Journal on Wireless Communications and Networking*, vol. 2009, pp. 1–15, 2009.