Two-level downlink scheduling for real-time multimedia services in LTE networks

Giuseppe Piro, *Student Member, IEEE,* Luigi Alfredo Grieco, *Member, IEEE,*
Gennaro Boggia, *Senior Member, IEEE,* Rossella Fortuna, and Pietro Camarda

Abstract

Long Term Evolution represents an emerging technology that promises a broadband and ubiquitous Internet access. But several aspects have to be considered for providing effective multimedia services to mobile users. In particular, in this work, we consider the design of a QoS aware packet scheduler for real-time downlink communications. To this aim, a novel two-level scheduling algorithm is conceived. The upper level exploits an innovative approach based on discrete-time linear control theory. Instead, at the lower level a proportional fair scheduler has been properly tailored to our purposes. The performance and the complexity of the proposed scheme have been evaluated both theoretically and by using simulations. A comparison with recently proposed scheduling strategies has been also presented, considering several network conditions and real-time multimedia flows. Particular attention has been devoted to the evaluation of the Quality of Experience provided to end users. Results have clearly shown that the proposed approach is able to greatly outperform the existing ones especially in the presence of real-time video flows.

Index Terms

LTE, Cellular Networks, Downlink Scheduling, QoS, QoE.

Authors are with the "DEE - Dip. di Elettrotecnica ed Elettronica", Politecnico di Bari, v. Orabona, 4 - 70125, Bari, Italy. e-mail: {g.piro,a.grieco,g.boggia,r.fortuna,camarda}@poliba.it.

A preliminary version of this work appeared as A Two-level Scheduling Algorithm for QoS Support in the Downlink of LTE Cellular Networks in Proc. of European Wireless, EW2010, Lucca, Italy, Apr. 2010.

Two-level downlink scheduling for real-time multimedia services in LTE networks

I. Introduction

The continuous raise of real-time multimedia services in the Internet and the need for an ubiquitous access to them are driving the evolution of cellular networks. Beside the huge bandwidth requirements, real-time multimedia flows need to be treated differently from other ones in order to reach a target Quality of Service (QoS). To face this ever growing demand for packet-based mobile broadband systems, the 3GPP [1] has introduced the LTE (Long Term Evolution) specifications as the next step of the current 3.5G mobile networks. In particular, an enhanced access network (i.e., the E-UTRAN, Evolved-UMTS Terrestrial Radio Access Network) and an evolved core network have been defined. At the present, more than 20 cellular operators worldwide have already stated a commitment to LTE (they represent together more than 1.8 billion of the 3.5 billion mobile subscribers in the world) and more than 32 million of LTE subscribers are forecast by 2013 [1]. Starting from this premise, it is clear that the optimization of all LTE aspects is a topic worth of investigation for both industry and academia communities, particularly considering multimedia applications.

In general, the most important objective of a multimedia service is the satisfaction of end users, i.e., the Quality of Experience (QoE). This is strictly related to the system ability to provide to application flows a suitable QoS [2], generally defined in terms of network delivery capacity and resource availability, i.e., limited packet loss ratio and delay. As example, a limited packet loss ratio enhances the quality of a reconstructed video, limiting distortions due to lack of video data packets, while a low delay allows to reproduce multimedia content at receiver side in real-time, i.e., with a small *playout delay*. In real-time multimedia services, such as VoIP or video-conference, end-to-end delay constraints in content delivery have to match the requirements related to the human perception of interactivity. For the Internet telephony, a delay of 100 ms is considered as the limit for a good perceived quality, while the delay has to be less than 300 ms for a satisfactory quality [3]. In order to respect audio/video synchronization, also for video delivery, the delay bounds have to be the same. In [4], for example, a delay of 200 ms is considered for video interactive applications. Once the video decoding process starts with a playout delay chosen

[1] http://www.3gpp.org

in this range (let us say [0 ms; 300 ms]), the respect of this deadline becomes mandatory for every encoded packet. In fact, every packet will be decoded with a playout delay after its generation time, and, if the packet does not arrive within the deadline, it will be considered lost. In this sense, in multimedia services, granting bounded delivery delays actually means lowering packet losses. This problem becomes very relevant for wireless access networks, such as LTE, due to the unpredictable behavior of radio links.

To face with QoE related issues, LTE specifications introduce a bearer identifying each flow that requires a particular policy [5]. The scheduler classifies packets belonging to a given bearer (i.e., a specified flow with QoS requirements) using a packet filter based on the well-known five-tuple: source and destination IP addresses, source and destination ports, protocol identifier. Thus, packet schedulers can allocate radio resources on a per-flow basis. To indicate the QoS level expected by the considered data flow, each bearer is associated to a QoS Class Identifier, defined by specifying up to four parameters: service class, priority, target delay, and packet loss ratio.

LTE specifications do not impose the adoption of any specific scheduler, thus leaving vendors free to implement their own solution. Anyway, despite the number of proposal already available in literature (see Sec. III for a summary of related works), lightweight algorithms able to schedule resource blocks for satisfying very sharp delay bounds still have to come. To bridge this gap, in this work we propose a novel approach to the problem, based on a two-layer scheduler, for real-time downlink communications.

Following LTE specifications, in our approach, time is seen as an endless sequence of frames, which are further split in time intervals. At the highest level, an innovative low complexity resource allocation algorithm has been designed using discrete time linear control theory (which will be referred to as Frame Level Scheduler, FLS). At the beginning of each frame, FLS computes the amount of data that each real-time source should transmit within the frame, to satisfy its delay constraint. Then, the lowest level scheduler, to ensure a good level of fairness among multimedia flows, assigns radio resources according to the Proportional Fair (PF) algorithm [6] subject to the constraint imposed by FLS. Radio resources left free by real time flows can be used to provide a best effort service using the PF algorithm [6], which enforces fairness also for this kind of flows.

The performance of the proposed algorithm has been tested using an open source simulator for LTE networks, i.e., the LTE-Sim [7]. A comparison with respect to recently proposed scheduling strategies, such as *Exponential* (*EXP*) *rule* and *Logarithmic* (*LOG*) *rule* [8] has been proposed, considering several network conditions and real-time voice/video multimedia flows. Particular attention has been devoted to the evaluation of the QoE perceived by end users. Results clearly show that the proposed approach is able to greatly outperform other schemes in terms of network performance and QoE, especially in the

presence of real-time video flows.

The rest of the paper is organized as follows: in Sec. II a basic background on the LTE technology is provided; Sec. III summarizes related works; in Sec. IV the two-level scheduling algorithm is designed; Sec. V reports simulation results; and finally the last section draws the conclusions and future research.

II. Overview of LTE

The requirements of LTE networks are very ambitious [6]: they will provide high peak data rates (up to 100 Mbps in downlink and 50 Mbps in uplink with 20 Mhz of bandwidth), increased cell edge throughput, less than 5 ms user plane latency, significant reduction of control plane latency, support for high user mobility, scalable bandwidth from 1.25 to 20 MHz, and enhanced support for end-to-end QoS. To fulfill these goals, the Radio Resource Management block has been designed to support a mix of advanced MAC and Physical functionalities, like packet scheduling, link adaptation, and Hybrid ARQ.

At the physical layer, as many existing wireless broadband systems, the LTE radio interface supports several duplexing techniques, based on frequency and time divisions. Radio transmissions are based on the Orthogonal Frequency Division Multiplexing (OFDM) modulation scheme. In particular, the Single Carrier Frequency Division Multiple Access (SC-FDMA) and the OFDM Access (OFDMA) are used in uplink and downlink transmissions, respectively. Differently from basic OFDM, they allow multiple access by assigning sets of sub-carriers to each individual user. OFDMA can exploit subsets of sub-carriers distributed inside the entire spectrum whereas SC-FDMA can use only adjacent sub-carriers. OFDMA is able to provide high scalability, simple equalization, and high robustness against the time-frequency selective radio channel fading. On the other hand, SC-FDMA is used in the LTE uplink to increase the power efficiency of user equipments (UEs), which are battery supplied. In addition, MIMO techniques can be exploited (both in downlink and uplink) to improve transmission reliability and data rate: it is possible to use up to a maximum of four transmission (receive) antennas [9].

According to [10], radio resources are allocated in a time/frequency domain. In the time domain, they are distributed every Transmission Time Interval (TTI), each one lasting 1 ms. Furthermore, each TTI is composed by two time slots of 0.5 ms, corresponding to 7 OFDM symbols in the default configuration with short cyclic prefix; 10 consecutive TTIs form the LTE Frame lasting 10 ms. In the frequency domain, instead, the whole bandwidth is divided into 180 kHz sub-channels, corresponding to 12 consecutive and equally spaced sub-carriers. A time/frequency radio resource, spanning over one time slot lasting 0.5 ms in the time domain and over one sub-channel in the frequency domain, is called Resource Block (RB) and corresponds to the smallest radio resource that can be assigned to a UE for data transmission. Note

that, given that the sub-channel dimension is fixed, the number of sub-channels varies accordingly to different system bandwidth configurations (e.g., 25 and 50 RBs for system bandwidths of 5 and 10 MHz, respectively).

At the base station, i.e., the so called *evolved node B* (eNodeB) the packet scheduler distributes radio resources among users. Scheduling decisions are strictly related to the channel quality experienced by UEs. In particular, the UE periodically measures this channel quality using reference symbols; then, it sends a Channel Quality Indicator (CQI) feedback to the eNodeB, with an uplink control messages [6]. The information about the quality of the time and frequency variant channel is exploited by the link adaptation module to select, for each UE, the most suitable modulation scheme and coding rate at the physical level with the objective of the spectral efficiency maximization. This approach is known as *Adaptive Modulation and Coding* (AMC) and it has been adopted by several wireless technologies, such as EDGE [11] and WiMAX [12]. Considering that each modulation scheme (i.e., QPSK, 16-QAM, and 64-QAM in LTE) corresponds to a fixed physical data rate, the link adaptation module establishes the maximum available physical data rate for each UE (based on the received channel quality information) for providing an optimal resource allocation among all users.

III. Related work

In LTE networks, the role of resource scheduling is very important because a great performance gain can be achieved by properly adapting the amount of frequency channels assigned to users in each TTI. As a consequence, the problem of finding low complexity algorithms able to distribute time slots and frequency carriers to users has attracted the attention of many researchers of the field. Such algorithms should take into account the expected QoS level, the behavior of data sources, and the channel status. The problem becomes even more challenging in the presence of users with different requirements in term of bandwidth, tolerance to delay, and reliability.

Classical approaches based on Maximum Throughput (MT), PF [13]-[16], Weighted Round Robin [17], and Adaptive Token Bucket [18] are not strictly applicable to handle real-time multimedia services. In fact, it is difficult to demonstrate their ability to satisfy strong requirements on packet loss ratio (PLR) and delay in a general network setting.

For this reason, several recent contributions (e.g., [8],[19]-[22]) propose channel-aware schemes that privilege flows having head-of-line packets approaching a target deadline. They mainly differ to each other depending on the weighting functions adopted to optimize fairness in bandwidth allocation and timeliness in packet delivery (see Tab. I for a complete description of parameters used by each scheduler).

TABLE I

PRIOR WORKS: PARAMETERS USED BY EACH SCHEDULERS

Scheduling Strategy	**SINR**	**Throughput**	**Head-of-Line Packet Delay**	**Target Delay**	**Target PLR**	**Queue Length**	**End-User Buffer Status**
MT [13]	**x**						
PF [13]	**x**	**x**					
QoS Oriented Time and Frequency Domain Packet Scheduler [14]	**x**	**x**					
CABA [15]	**x**	**x**					**x**
Proportional Fair Multiuser Scheduling [16]	**x**	**x**					
ATBFQ [18]	**x**	**x**					
Quality-Driven Cross-Layer Scheduler [17]	**x**			**x**			
Packet Scheduling Scheme to Support Real-Time Traffic [19]	**x**			**x**		**x**	
Frequency-Time Scheduling for Streaming Services [20]	**x**	**x**	**x**	**x**			
Multi-Service QoS Guaranteed Based Downlink Cross-Layer Scheduler [21]	**x**	**x**	**x**	**x**	**x**		
EXP-PF [22]	**x**	**x**	**x**	**x**	**x**		
M-LWDF [22]	**x**	**x**	**x**	**x**	**x**		
EXP Rule [8]	**x**	**x**	**x**	**x**			
LOG Rule [8]	**x**	**x**	**x**	**x**			

Notably, in [8], a very thorough discussion on related works in this field has been reported. Furthermore, *EXP* and *LOG rules* have been presented as the most promising approaches for downlink scheduling in LTE systems with delay-sensitive applications.

We remark that all the aforementioned approaches cannot offer any strict guarantees on packet delivery delay, which, for the considerations already reported in Sec. I, plays a major role in the end-user satisfaction.

In general, to meet QoS constraints, it is not sufficient to provide guarantees on the average packet delay, but it is necessary to enforce guarantees on the upper bound of packet delays. It is worth to note that only in [23] a generic scheduling for channel-adaptive wireless networks has been proposed to provide absolute delay guarantees to real-time flows, but this scheme presents a very high computational complexity and it is hard to adopt it in the LTE radio interface.

Furthermore, another important limitation of discussed prior works is that they do not consider that some flows with pending data near to the deadline could experience a sudden decrease of the channel quality just before packet deadline expiration, with the consequent violation of the target delay.

Finally, all previous works consider the problem of radio resource allocation TTI by TTI without any medium-term planning; such a planning could improve the system performance by exploiting the requirements of all competing flows over a larger time-scale.

For this reason, herein we propose a two-level scheduler able to target bounded packet delays by taking into account the advantage of statistical multiplexing in resource allocation.

To demonstrate the effectiveness of our approach with respect to the best solutions proposed so far, we chose to compare our proposed scheduler with both *EXP rule* and *LOG rule*, considering several network conditions, scenarios with real-time multimedia flows and best effort flows, and analyzing also the impact of the PLR on the provided QoE. Results will emphasize the effectiveness of the proposed allocation scheme describing how it is able to fully respect QoS requirements of multimedia flows with respect to other ones.

IV. Two-level scheduling

The conceived novel scheduling strategy targets real-time service provisioning in the LTE downlink. It has been built on two distinct levels (see Fig. 1) that interact together in order to dynamically assign radio resources to UE. They take into account the channel state, the data source behaviors, and the maximum tolerable delays.

At the highest level, an innovative resource allocation algorithm, namely FLS, defines frame by frame the amount of data that each real time source should transmit to satisfy its delay constraint. To solve the problem using a solution with a low computational complexity, FLS exploits a discrete-time linear control loop [24]. Once FLS has accomplished its task, the lowest layer scheduler, every TTI, assigns RBs using the PF algorithm [6] by considering bandwidth requirements of FLS.

In other words, FLS defines on the long run (i.e., in a single frame) how much data should be transmitted by each data source. The lowest layer scheduler, instead, allocates resource blocks in each

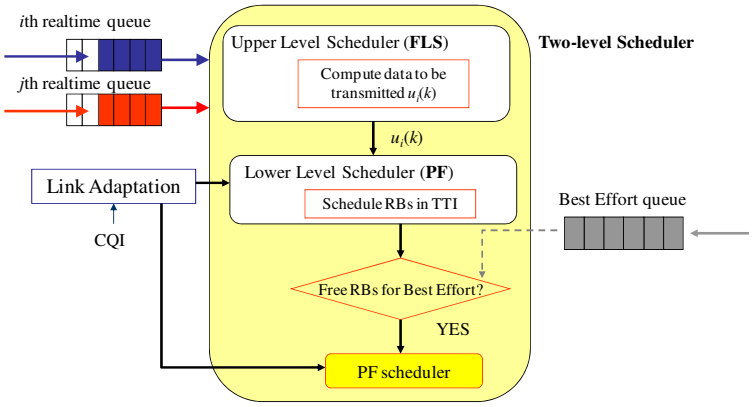


Fig. 1. The two-level scheduling algorithm.

TTI to achieve a trade-off between fairness and system throughput. It is important to note that FLS does not take into account the channel status. On the contrary, the lowest layer scheduler assigns RBs first to flows hosted by UEs experiencing the best channel quality and then (i.e., when these flows have transmitted the amount of data imposed by FLS) it considers the remaining ones. In particular, the lowest layer scheduler decides the number of TTIs/RBs (and their position in the time/frequency domains) in which each real-time source will actually transmit its packets.

It is very important to remark that the proposed approach is very general and it is independent on the model used for describing incoming data. For this reason, we do not need stochastic flow models. In fact, the control theoretic approach describes a flow as a signal modelling the bit-rate produced by the application layer.

A. The upper level of the scheduler

The FLS scheduler (that is, the upper level of our two-level scheduler) has been designed using discrete-time linear control theory arguments. In our system, we suppose that N active traffic flows share the wireless channel. Packets waiting for transmission are stored in a queue associated to each flow. FLS evaluates the transmission needs of all queues at the beginning of each LTE frame.

The role of the FLS scheduler at the upper level is to evaluate, by a closed control loop scheme (see Sec. IV-A), the quota of data, $u_i(k)$, that the i-th real-time source should transmit in the k-th frame to meet its QoS constraints. A control law is used to compute $u_i(k)$ and it is defined to provide sharp delay bounds to real time flows as will be shown below.

The quota of data $u_i(k)$ assigned to the generic i-th flow is decoupled from the ones assigned to

the remaining flows. For this reason, we describe the mathematical analysis of the proposed scheduling strategy considering only a generic flow (i.e., the i-th flow). However, the same considerations are valid for all flows.

Let be $t_{k,i}$ the starting time of the k-th frame for the i-th real-time source. Therefore, the sampling interval $\Delta t(k) = t_{k+1,i} - t_{k,i}$ is equal to the LTE frame duration, T_f.

Now, the following equation holds:

$$q_i(k+1) - q_i(k) = d_i(k) - u_i(k), \tag{1}$$

where $q_i(k)$ is the i-th queue length at time $t_{k,i}$; $q_i(k+1)$ is the i-th queue length at time $t_{k+1,i}$; $u_i(k)$ corresponds to the amount of data that is transmitted during the k-th frame; $d_i(k)$ is the amount of data that filled the queue during the k-th frame, i.e., it models the behavior of the data source feeding the i-th queue.

In the following, we will refer to $Q_i(z)$, $D_i(z)$, and $U_i(z)$ as the $\mathcal{Z}$-transforms of the signals $q_i(k)$, $d_i(k)$, and $u_i(k)$, respectively.

At the beginning of the k-th frame, FLS has to compute the quota of data $u_i(k)$ that the i-th flow should transmit in the considered frame. Thus, we have to design a control law that should provide bounded packet delays and, at the same time, the BIBO (i.e., Bounded Input Bounded Output) stability [24] of the system defined by eq. (1).

We start assuming the following general control law:

$$u_i(k) = h_i(k) * q_i(k) \tag{2}$$

where the '$*$' operator is the discrete time convolution [24].

Eq. (2) means that the amount of data to be transmitted by the i-th flow during the k-th LTE frame is obtained by filtering the signal $q_i(k)$ (i.e., the queue level) through a time-invariant linear filter with pulse response $h_i(k)$ or, equivalently, with the transfer function $H_i(z) = \mathcal{Z}[h_i(k)]$ [24].

Combining eqs. (1) and (2), we obtain that our scheduling algorithm realizes the control loop shown in Fig. 2, with the set point $q_i^T = 0$. This means that our control algorithm tries to target empty queues using a linear regulator with transfer function $H_i(z)$.

From now on, the pulse response of the system will be referred to as $h_{s_i}(k)$, so that the following equality holds:

$$q_i(k) = h_{s_i}(k) * d_i(k) \ . \tag{3}$$

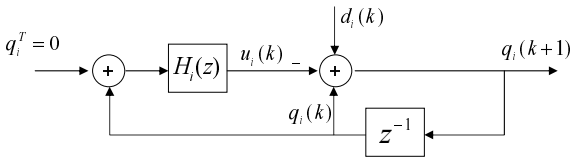


Fig. 2. Control loop of the allocation algorithm at the upper layer of the scheduler.

Assuming $q_i(0) = 0$ (i.e., empty queues at the beginning), our design strategy is to find the proper function $H_i(z)$ that ensures the BIBO (Bounded Input Bounded Output) stability to the system and guaranteed queuing delays.

As well known, BIBO stability means that the output of the system remains bounded in amplitude, provided that the input is bounded too. This property (which is equivalent to the asymptotic stability in linear systems) ensures that FLS will never try to allocate an infinite bandwidth because the system input (i.e., the incoming data rate) is bounded in amplitude since any realistic application cannot produce an infinite packet rate. BIBO stability is only a propaedeutic requirement because it is not able to guarantee any specific performance index related to transient and steady-state system behavior [24]. Hence, in our case, FLS is designed to be BIBO-stable and, at the same time, to satisfy delay bounds of served flows.

To this aim, we need the following theorem (see Appendix A for the proof).

Theorem 1: The system is BIBO stable and a queuing delay for the i-th queue smaller than $M_i + 1$ sampling intervals can be provided, if the closed-loop response to the Kronecker pulse $\delta(k)$ [24] (i.e., the system pulse response) has the following expression:

$$h_{s_i}(k) = \sum_{n=0}^{M_i} c_i(n)\delta(k-n) \tag{4}$$

where M_i is the length of the pulse response and

$$0 \leq c_i(n) \leq 1 \ \forall n; \qquad c_i(n) \geq c_i(n+1), n \geq 1 \quad \text{with } c_i(n) \in \mathbb{R}. \tag{5}$$

Remark 1: Considering that each sampling interval lasts T_f, the upper bound of the queuing delay, τ_i, is:

$$\tau_i = (M_i + 1)T_f \ . \tag{6}$$

Now we can design the transfer function $H_i(z)$. In fact, the eq. (4) for the system pulse response is satisfied when the transfer function of the controller is (see Appendix B):

$$H_i(z)=\frac{U_i(z)}{Q_i(z)}=\left[(1-z)\sum_{n=0}^{M_i}c_i(n)z^{-n}+1\right]\Bigg/\sum_{n=0}^{M_i}c_i(n)z^{-n}. \tag{7}$$

Thus, FLS is able to grant bounded packet delays.

It is very important to remark that the proposed approach is able to guarantee the same bounded delay also in the case of a temporary channel disturbance, where a flow cannot transmit the quota of data $u_i(k)$ computed by FLS until the end of the current frame. In fact, the amount of data scheduled in every T_f, depends on the amount of data present in the transmission queue. If a sudden reduction of the channel quality in the current temporal slot stops the transmission of the scheduled data, FLS will take into account the pending packets, scheduling them in the next frame and defining a greater amount of data to be transmitted. Thus, it will drain the transmission queue, avoiding expirations of packet deadlines.

1) FLS computational complexity: The FLS algorithm can be easily casted in the LTE downlink scheduler because each eNodeB knows the transmission queue of all active flows in downlink (i.e., q_i, $1 \leq i \leq N$). As a consequence, it can easily compute the u_i values, with $1 \leq i \leq N$, at the beginning of each frame.

To evaluate the computational complexity of FLS, we can express the control low by considering eq. (7) in the time domain:

$$u_i(k) = q_i(k) + \sum_{n=2}^{M_i}\left[q_i(k-n+1)-q_i(k-n+2)-u_i(k-n+1)\right]c_i(n) \tag{8}$$

that can be obtained .

It is clear that, for each flow, the computation of $u_i(k)$ requires (M_i-1) multiplications and $3(M_i-1)+1$ sums, that is the computational complexity for each flow is $O(M_i)$. As a consequence, if in the E-UTRAN system there are N active downlink real time flows, the total computational complexity is $O(NM^*)$ where $M^* = \max_i\{M_i\}$ with $i = 1, \ldots, N$.

2) The role of coefficients c_i : To provide a more intuitive explanation of the proposed approach, we clarify the impact of coefficients c_i. FLS has been conceived as a discrete-time control loop scheme which defines the quota of data that each flow should transmit every LTE frame in order to obtain bounded packet delays. In particular, at the beginning of every frame, FLS computes the quota of data that the i-th flow should transmit during the frame using Eq. (8). c_i coefficients are used in the discrete-time filter modeling our system. Such a filter has as input variable the incoming data packets and as output variable the queue length, as shown in Fig. 2 and formalized in Eqs. (2) and (3). The fact that this filter has a

finite pulse response with M_i coefficients means that FLS is able to guarantee bounded delays, because the transmission of enqueued data is spread over no more than M_i consecutive sampling intervals. This is the meaning of the afore-presented Theorem 1 and of the consequent remarks. The values of coefficients c_i determine how the enqueued data are spread over the M_i consecutive sampling interval, as pictured in Fig. 12 of Appendix A.

3) Impact of packet retransmission to the scheduling behavior: As in a generic wireless networks, also for LTE radio interface there is the probability to have errors at the PHY layer. LTE provides a couple of retransmission techniques (the first one at the MAC layer and the second one at the RLC layer) which are used for recovering part of data lost at the PHY layer. We discuss now how the ARQ process impacts on the FLS behavior. FLS works on top of ARQ so that, retransmissions are simply perceived as a reduction of the available bandwidth which is mitigated by the closed-loop. In other words, ARQ impacts on the transmission queue level $q_i(k)$, which is in turn used by FLS to throttle the bandwidth assignment $u_i(k)$. In this way, the behavior of lower layers is compensated by FLS at each new frame by new bandwidth assignments that take into account what happened in past frames.

B. The lower level of the scheduler

For each of the ten TTIs forming a frame, the lower level scheduler allocates the RBs to real time flows. At the k-th TTI, only flows that have not yet transmitted their quota $u_i(k)$ in the previous TTIs of the same frame will be scheduled. Thus, as soon as a real time source has transmitted its quota of data $u_i(k)$ defined by FLS, it looses the opportunity to transmit until the beginning of next frame.

To achieve a high level of fairness among multimedia flows, the lower layer scheduler uses the PF algorithm [6]. In particular, every TTI, the link adaptation module evaluates the maximum instantaneous supportable data rate for each UE, in each sub channel. This value is computed using feedbacks on channel quality sent by the UE in the previous TTI (see Sec. I). The PF algorithm assigns RBs to downlink connections belonging to UEs with the best ratio w, computed as the instantaneous available data rate over the average data rate. That is, with reference to the i-th UE in the j-th sub-channel:

$$w_{i,j} = R^M_{i,j}/\overline{R}_i, \tag{9}$$

where $R^M_{i,j}$ and $\overline{R}_i$ are the instantaneous maximum available data rate and the estimated average data rate, respectively.

We note that the $\overline{R}_{i,j}$ value is updated every TTI using a weighted moving average formula and by

taking into account the effective quota of data

$$\overline{R}_i(k) = 0.8\overline{R}_i(k-1) + 0.2R_i(k-1), \tag{10}$$

where $R_i(k-1)$ represents the data rate achieved by the i-th users during the previous sub frame.

It is worth to note that, for a given amount of data $u_i(k)$, the number of RBs required to transmit is not fixed, but it depends on both the digital modulation scheme chosen by link adaptation every TTI and the protocol/physical overhead.

For what concern the radio resources left free by real time flows, they are assigned using even the PF scheduler in order to provide a high degree of fairness.

Fig. 3 shows a simple example of the proposed resource allocation scheme: the RBs can be assigned to best effort flows if and only if all real time flows have been served according to FLS rules.

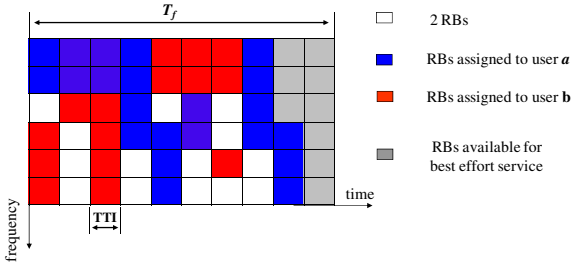


Fig. 3. Example of the resource allocation scheme.

The proposed scheduler has been conceived using a two-layer approach. The upper layer (i.e., the FLS scheduler) has been designed for guaranteeing bounded packet delays to multimedia flows. It is clear that FLS targets can be achieved only when the lower layer scheduler is able to assign to each multimedia flow the quota indicated by FLS itself (i.e., the quota of data $u_i(k)$). To this aim, the lower layer strategy assigns to multimedia flows a higher priority than the one assigned to other kind of flows, until they have accomplished the transmission of $u_i(k)$ data ruled by FLS. An important concept we have to consider is that the amount of radio resources (i.e., the number of PHY resource blocks) that the i-th flow requires for sending $u_i(k)$ depends on the channel quality experienced by the UE that receives that flow. For this reason, FLS will be able to guarantee bounded delays if and only if the channel quality of each UE receiving multimedia flow is large enough to accommodate FLS assignments. If a given user perceives a bad channel condition, neither FLS nor any other algorithms would be able to guaranteed targets for which they have been designed. But, in this case, the proposed approach has an advantage with respect to other scheduling algorithms due to the use of a control loop scheme. In fact, if a user measures a

reduction of the channel quality in a given temporal slot and it is not possible to transmit the whole $u_i(k)$ assignment, FLS algorithm will implicitly take into account the pending packets (which remain stored in the transmission queue), thus scheduling them in the next frame.

V. Performance Evaluation

To study the effectiveness of the proposed scheduler, we used the LTE-Sim [7], an open source simulator for LTE networks. A comparison with the well-known scheduling strategies *LOG rule* and *EXP rule* [25] has been also provided. Furthermore, to appreciate the effectiveness of our proposed allocation scheme in realistic settings, both the influence of inter-cell interference and the impact on the QoE perceived by end users for real-time flows have been analyzed.

Simulation results demonstrate that the proposed resource allocation scheme is able to respect target delays of real time flows (i.e., the considered QoS constraints), in all operative conditions, assuring the best QoE with respect to other scheduling strategies.

Finally, we described as the proposed approach, despite its simplicity, it is able to provide better performance for multimedia flows.

A. Simulation Scenario

We have developed a realistic multi cell scenario composed by 19 cells with radius equal to 0.5 km. To guarantee for each cell a bandwidth of 10 MHz in the downlink, frequencies of the first operative LTE bandwidth[2] are distributed among clusters composed by 4 cells (see Fig. 4). In each cell, there are one eNodeB and a variable number of UEs in the range [10-20]. Mobility of each UE traveling cells is described with the random way-point model [27]. Moreover, to analyze both pedestrian and vehicular users, we have considered a speed equal to 3 and 120 km/h, respectively.

We have imposed that each UE receives at the same time one video flow, one VoIP flow, and one best effort flow, as shown in Fig. 5.

For the video flow, herein we consider the results obtained using a traffic trace created from the video test sequence “highway.yuv”. Note that we tested performance of the scheduling algorithms also with other video clips (i.e., “mobile.yuv” and “foreman.yuv”), achieving very similar results not reported here for lack of space[3]. The original sequence at 25 frame per second [fps], CIF resolution 352×288 and

[2]The first operative bandwidth for LTE is defined in the range [1929-1980] MHz for the uplink and [2110-2170] MHz for the downlink, in FDD mode [26].

[3]All the considered video sequences are available at http://www.hlevkin.com/.

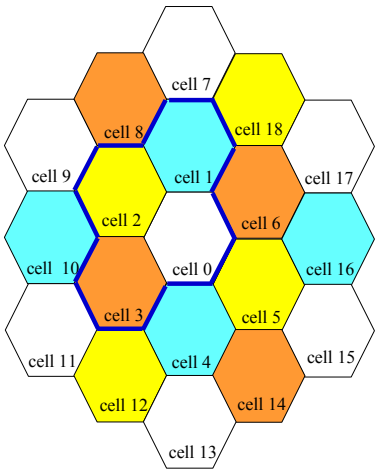


Fig. 4. Simulation scenario with 19 cells and a cluster composed by 4 cells.

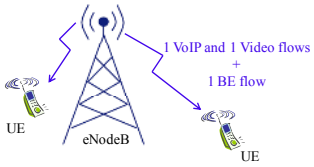


Fig. 5. LTE simulated scenario.

YUV format has been firstly repeated for the whole simulation time. Then, the obtained video sequence has been compressed using H.264 standard compression at the average coding rate of 128 kbps.

Instead, for G.729 voice flows we have adopted an ON/OFF Markov model, where the ON period is exponentially distributed with mean value 3 s, and the OFF period has a truncated exponential pdf with an upper limit of 6.9 s and an average value of 3 s [28]. During the ON period, the source sends 20 bytes sized packets every 20 ms (i.e., the source data rate is 8 kbps), while during the OFF period the rate is zero because we assume the presence of a voice activity detector.

Finally, for the best effort flows we have considered infinite buffer sources.

Each simulation lasts 100 s and all simulation results are averaged over 5 simulations.

B. System model for LTE E-UTRAN

The main simulation parameters used in the LTE-Sim simulator are summarized in Tab.II.

TABLE II

SIMULATION PARAMETERS.

Parameter	Value
Simulation length	100 s
Physical Detail	Carrier Frequency: 2 GHz; Bandwidth for the DL: 5 MHz; Symbol for TTI: 14; SubFrame length: 1 ms; SubCarries per RB: 12; SubCarrier spacing: 15 kHz; eNodeB: Power transmission=43 dBm equally distributed among sub-channels; 2 antenna ports; Modulation Scheme: QPSK, 16QAM, and 64QAM with all available coding rates target BLER: 10%
Overhead	RTP/UDP/IP with ROCH compression: 3 bytes MAC and RLC: 5 bytes; PDCP: 2 bytes; CRC: 3 bytes L1/L2: 3 symbols
Cell layout	radius: 0.5 km
RLC ARQ	activated with maximum 5 retransmissions
CQI	Full bandwidth and periodic reporting scheme. Measured period: 2 ms.
Number of UEs	10, 15, 20
Traffic Model	real time traffic type: H264, VoIP; best effort flows: infinite buffer

Packet scheduling analysis is based on the parameters for the downlink of the LTE E-UTRAN suggested by 3GPP specifications [9].

At the PHY layer, we suppose that the eNodeB uses two antenna ports and a power transmission equal to 43 dBm, uniformly spread over the all 50 available sub-channels. The Transport Block Size, or in other words, the quota of data that a flow can transmit at the MAC layer during a TTI and using one sub-channel, is obtained from the selected MCS, taking into the account the physical configuration proposed in [29]: normal prefix code, 2 antenna ports, 3 OFDM symbols for the Physical Downlink Control Channel (PDCCH), no sync signal, and Physical Broadcast Channel (PBCH) not present.

Regarding channel model, a propagation loss model for an urban cell has been considered according to [9]. Such a model takes into account four different phenomena affecting channel conditions: (i) the path loss, (ii) the penetration loss, (iii) the shadowing, and (iv) the effect of fast fading due to the signal multipath. In particular, the path loss, P_L, is given by the expression $P_L = 128.1 + 37.6 \log d$, where d is the distance between the eNodeB and the UE, in kilometers. The large scale shadowing fading has been modeled through a log-normal distribution with 0 mean and 8 dB of standard deviation. According to [9], the penetration loss has been set to 10 dB. Finally, the time-frequency correlated signal multipath

is modeled by using the Rayleigh fading channel model proposed in [30].

We chose to use a full bandwidth and periodic CQI reporting scheme. Each UE estimates every 2 TTIs the SINR (Signal to Interference plus Noise ratio) of the received downlink reference signals for all available downlink sub-channels. Then, it maps these values to a set of CQI feedbacks which will be forwarded to the eNodeB using the uplink channel. As described in [7], the CQI value is obtained as a quantized version of the estimated SINR in order to guarantee a BLER (BLock Error Rate) at least equal to 10%. The mapping between SINR and CQI is performed through the BLER-SINR curves [7].

After scheduling decisions, the packet scheduler selects a proper MCS to be used in the downlink for a given scheduled UE. The eNodeB assigns to that user a MCS by using the Exponential Effective SINR Mapping method [31]. In particular, the eNodeB firstly computes the effective SINR considering the latest CQI feedbacks received by the scheduled UE for the assigned sub channels. Then, it maps this value to a proper MCS to guarantee, for that SINR, a BLER equal to 10%.

In our simulation, PHY errors have been also modeled, as proposed in [7].

1) Implementation of schedulers: The first important scheduling parameter is the target delay for real-time multimedia flows. In real-time services, such as VoIP or video-conference, the maximum allowed end-to-end delay has to be chosen in the range 100÷200 ms for assuring the final users perception of interactivity [3]. Accordingly, target delay for the last hop (which is in our analysis the radio link between eNB and UE) should be less than the aforementioned values.

In our simulations we used a set of target delays (i.e., [40-100] ms), which is even broader than requirements suggested in 3GPP specifications [32]. We considered also delay constraints stronger than those proposed by 3GPP for demonstrating the effectiveness of the proposed approach (as well as LOG and EXP rules) under strenuous network conditions. It is important to note that we adopted the same target delays for both video and voice flows in order to ensure that they are synchronously played out at the same UE, which is of major importance in video-conferencing.

To provide a further insight, we remark that packets are deleted from the transmission queue only when they expire, i.e., when they are not transmitted within the deadline. This avoids bandwidth waste. In fact, in multimedia communications, out-to-date packets received by mobile station after their deadline have to be considered lost, because they are no more usable by the decoding process. In this sense, limiting the number of delayed packets, actually means a reduction of the PLR. This means also that the ability to respect target delays can be derived by examining the PLR, which increases when the scheduler is not able to timely serve real-time packets.

When the FLS has been used, we have considered in simulations $M_i = 3, 5, 7, 9$, according to the

TABLE III

EXP RULE AND LOG RULE SCHEDULING METRICS

Scheduler	**metric** m	**parameters**
EXP rule	$m_{i,j} = b_i exp(\frac{a_i w_i}{c+(\frac{1}{N}\sum_j a_i w_i)^\eta}) \times K_{i,j}$	$b_i = \frac{1}{E[K_i]}$, $c = 1$, $a_i = \frac{6}{d_i}$
LOG rule	$m_{i,j} = b_i log(c + a_i w_i) \times K_{i,k}$	$b_i = \frac{1}{E[K_i]}$, $c = 1.1$, $a_i = \frac{5}{d_i}$

chosen target delay. Moreover, with reference to eq. (4), we set the $c_i(n)$ coefficients as follows:

$$c_i(0) = 0; \quad c_i(n) = 1 - (n-1)/M_i \quad \forall n = 1, \ldots, M_i. \tag{11}$$

In this way $h_{S_i}(k)$, described in eq. (4), has a linear pulse shaping, so that enqueued data waiting for transmission will be spread uniformly over M_i consecutive sampling interval. (see Sec. IV-A2 for further details).

Finally, both *EXP rule* and *LOG rule* have been developed as proposed in [25]. For these schedulers, best-effort flows are managed by using the common PF algorithm (i.e., as described in Sec. IV-B).

Tab. III reports scheduling metric used for multimedia flows with the related parameter sets for both EXP rule and LOG rule. We note that i, j, $K_{i,j}$, and w_i represent the user identity, the sub channel identification, the spectral efficiency of the $i-th$ user for the $j-th$ sub channel, and the head of line packet delay for the $i-th$ user, respectively. These parameters have been optimized in [25] following guidelines proposed in [33].

C. Scheduling performance

The performance of *EXP rule*, *LOG rule*, and FLS have been evaluated by varying the number of UEs, the speed, and the target delay imposed to real time flows, also considering the inter-cell interference. The comparison has been divided on the basis of several performance indexes such as the PLR and the QoE of multimedia sessions as well as the goodput and the fairness of best effort flows.

Regarding multimedia flows, we note that PLR is a standard metric traditionally used to evaluate the Quality of Service offered by the system at network layer. Moreover, additional metrics for Quality of Experience evaluation have to be considered to generally evaluate system performance in terms of user satisfaction. Since an optimal metric for quality of experience evaluation has not yet been standardized, we choose the Peak Signal to Noise Ratio (PSNR) [34], which is nowadays one of the most diffused metrics for evaluating user satisfaction, together with the interactivity level, in real time video applications.

Since best-effort flows do not required strict QoS specifications, we chose to compare scheduling strategies by considering both aggregate goodput and fairness index provided to by these flows.

Figs. 6 and 7 show the PLR achieved for video and VoIP flows, respectively. It is possible to observe that the PLR increases with the number of UEs, due to the higher network load. We note also that users with the highest speed achieve the greatest PLR. When the user speed increases, in fact, channel quality changes in two consecutive sub frames are more likely; thus, there could be more frequent errors in MCS selection. As expected, a lower value of the target delay implies a higher value of PLR due to a larger quota of packets violating the deadline.

The most important result we have obtained is that the smallest PLR is obtained using the proposed allocation scheme. This is a clear demonstration that imposing a bound on the maximum tolerable delay can greatly improve the quality of multimedia services in LTE system. Moreover, it is very important to note that performance reached by *LOG rule* and *EXP rule* are different from those described in [25]. The reason is that we have analyzed scheduling performance in a more complete and complex scenario with respect to those simulated in [25] . Thus, in these considered environments *EXP rule* is able to obtain better performance with respect to *LOG rule*.

It is worth to note that VoIP flows experience significantly smaller PLRs than video ones. The reason is that VoIP flows, having a lower source bit rate, get the highest priority from the PF scheduler.

Finally, we verified that PLR is almost uniformly distributed over the UEs within the cell.

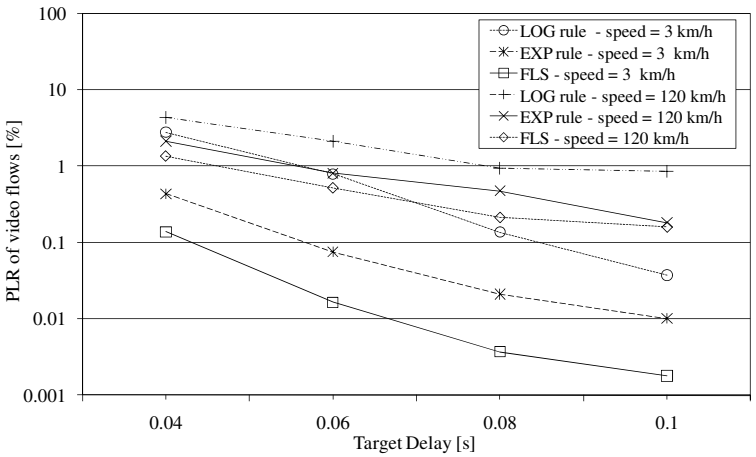
In Sec. V-D, it will be shown as a higher value of PLR translates in a bad level of QoE.

To study the behavior of best-effort flows, Fig. 8 shows the aggregate goodput, defined as the rate of useful bits successfully transmitted by this kind of flows during th whole simulation. We note that the goodput decreases as the number of UEs and the user speed increase. When the number of UEs is equal to 10, all the considered schedulers register a similar aggregate goodput. Moreover, when the number of UEs increases, note that, using both *LOG rule* and *EXP rule* allocation schemes, we obtain a higher goodput for best-effort flows with respect to the use of the FLS algorithm. This result was expected because, as seen in Figs. 6 and 7, *LOG rule* and *EXP rule* provide a worser service to multimedia flows (with respect to our proposal), thus leaving a higher quota of bandwidth for best-effort flows.

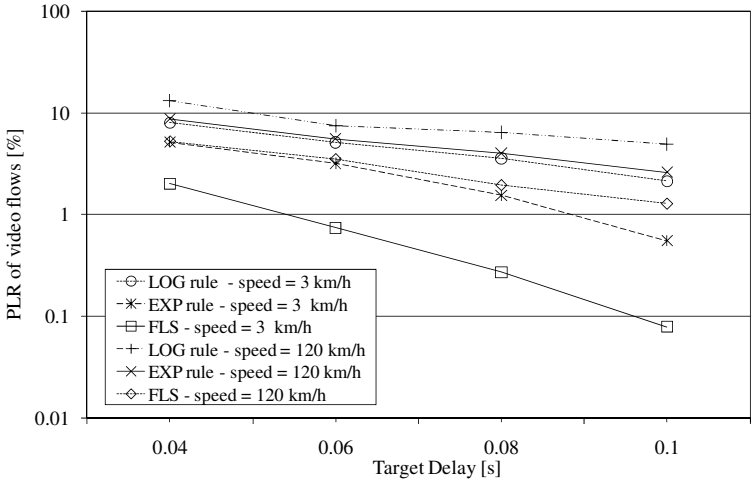
Finally, Fig. 9 shows that all considered schedulers provide a high degree of fairness.

D. QoE of multimedia flows

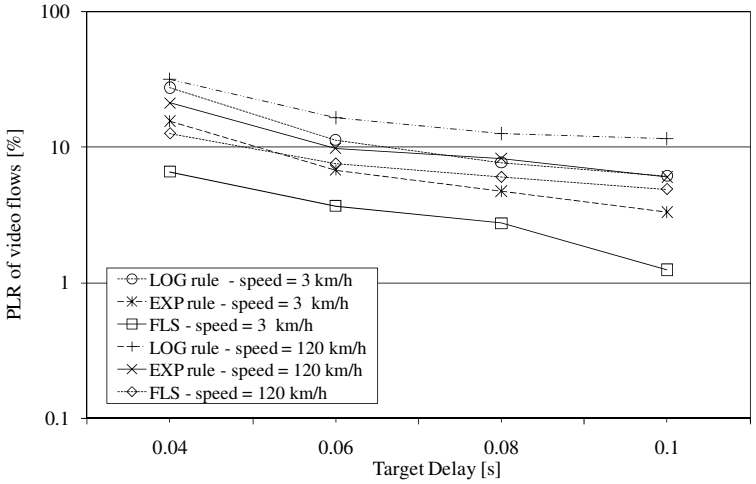
Latest considerations we would remark regard the impact of the PLR on the QoE perceived by users receiving VoIP and video flows. This analysis is very important because it allow us to understand how



(a)

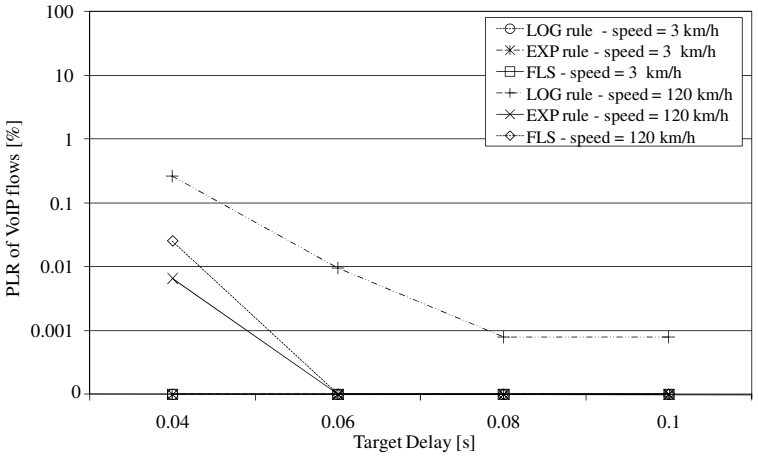


(b)

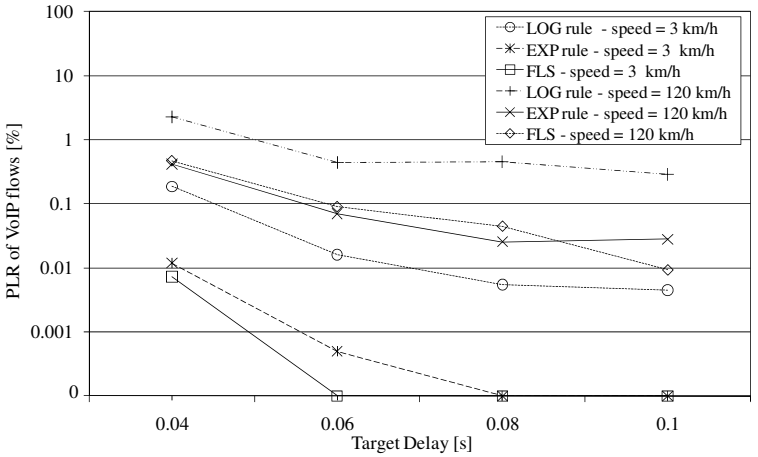


(c)

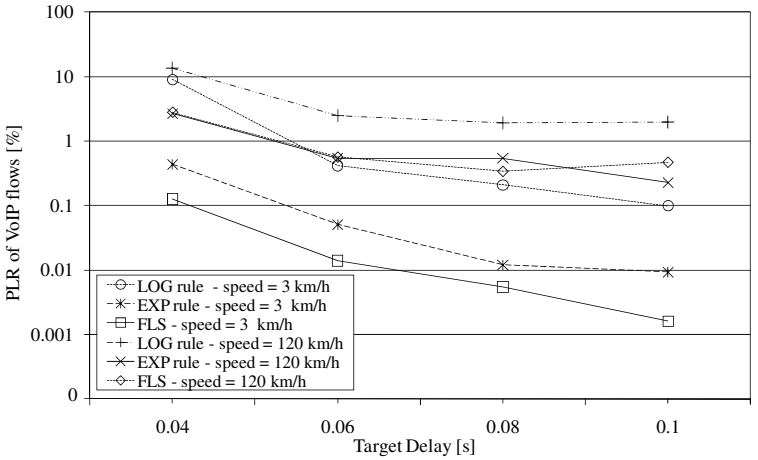
Fig. 6. Packet Loss Ratio (PLR) of Video flows in a scenario with (a) 10, (b) 15, and (c) 20 UEs.



(a)

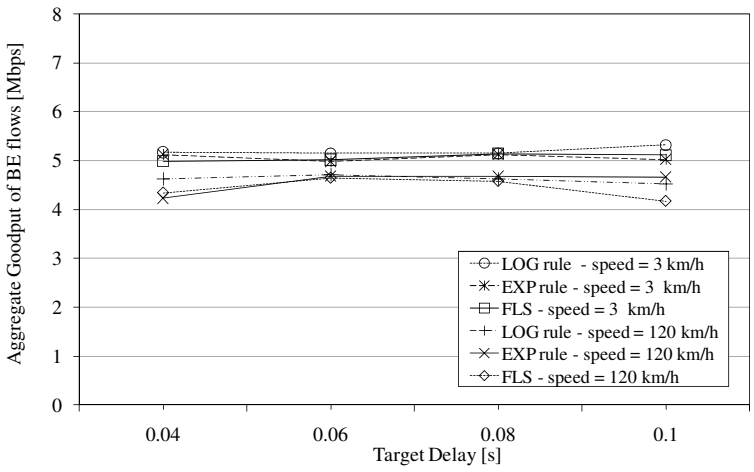


(b)

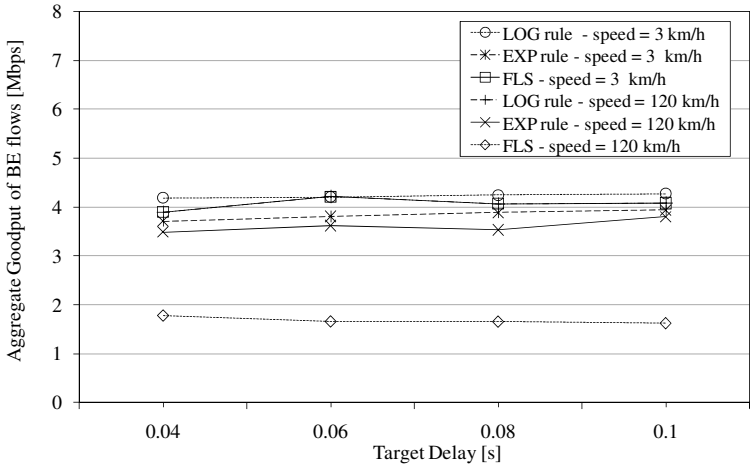


(c)

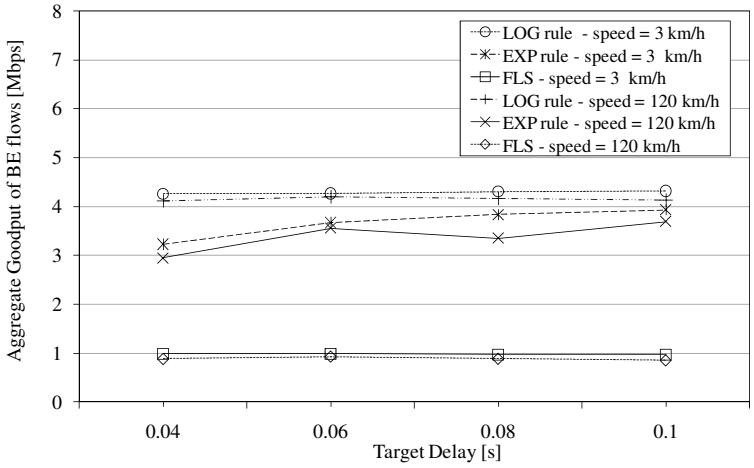
Fig. 7. Packet Loss Ratio (PLR) of VoIP flows in a scenario with (a) 10, (b) 15, and (c) 20 UEs.



(a)

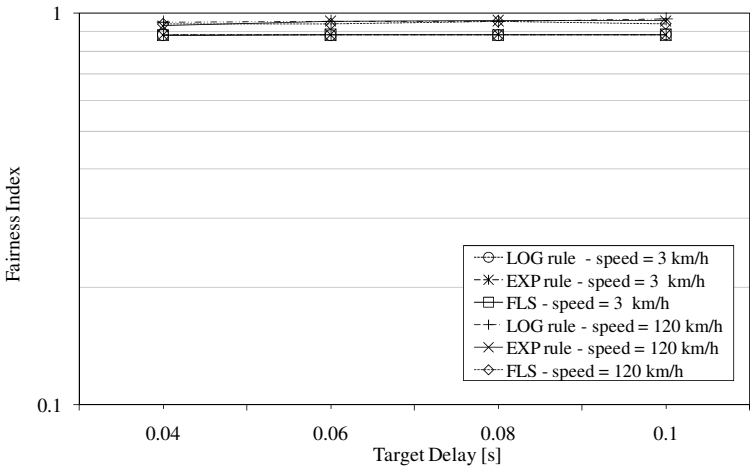


(b)

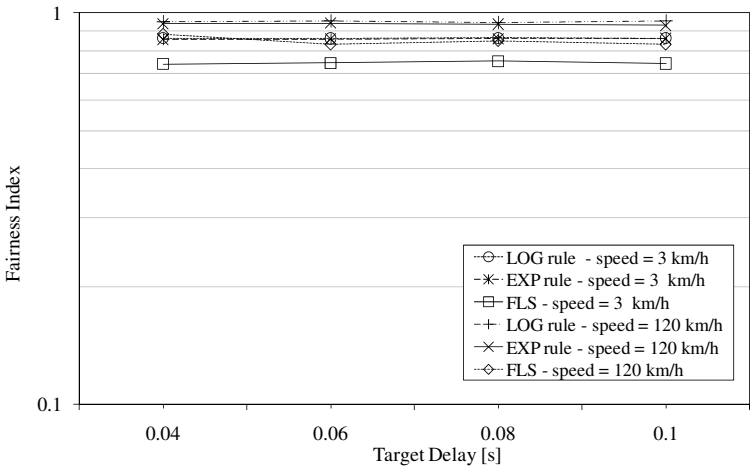


(c)

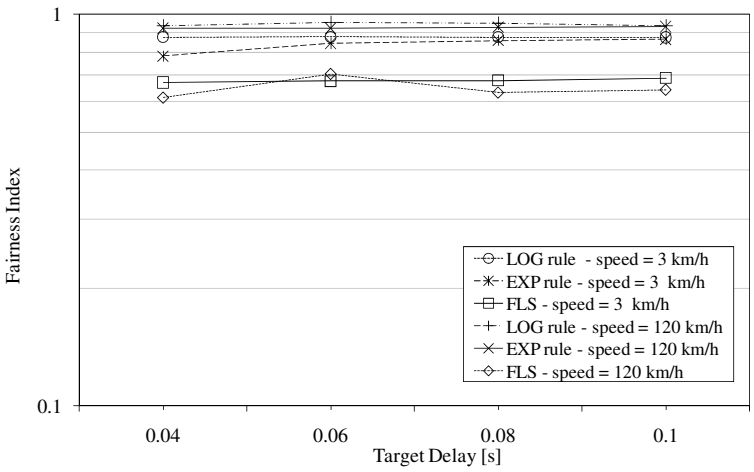
Fig. 8. Goodput of FTP flows in a scenario with (a) 10, (b) 15, and (c) 20 UEs.



(a)



(b)



(c)

Fig. 9. Fairness Index of Best Effor flows in a scenario with (a) 10, (b) 15, and (c) 20 UEs.

the packet loss influences the quality of the received real-time flow.

To estimate the perceived speech quality of VoIP flows we have used the well-known Mean Opinion Score (MOS) using the ITU E-model [35]. In particular, the output of the E-model (i.e., the transmission rating factor) has been computed assuming a constant end-to-end packet delay, equal to the target delay, due to the presence of the playout buffer at the receiver, as described in [35]. Then, the transmission rating factor has been mapped to the proper MOS value, considering the mapping function proposed in [35].

Imposing a playout buffer at the receiver equal to the target delay, we have obtained a MOS higher that 4 for all considered schedulers. According to [36], the MOS value above 3.6 corresponds to satisfaction for almost users. For this reason, all schedulers are able to provide a good speech quality in all operative conditions. This result was expected given the very small packet loss ratios provided to VoIP flows (see Fig. 7).

The quality of received video data has been estimated computing the PSNR between the transmitted and the received videos. PSNR [34] offers a way for quantifying the impact of losses on video quality and, together with the interactivity level assured by the system, can be considered one of the key metrics for Quality of Experience evaluation in real-time video streaming systems. The PSNR has been computed without considering the quality degradation inserted by the encoding process, in order to highlight the video distortion due to packet losses. [4]

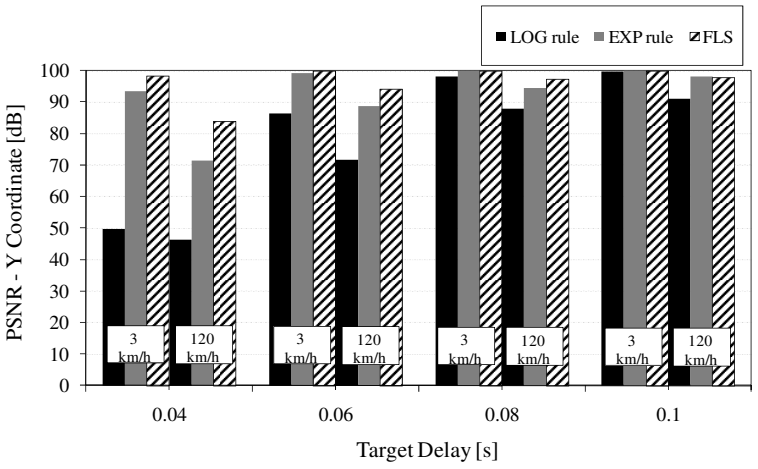
In particular, Fig. 10 shows the PSNR computed for the Y video components which has the strongest impact on the QoE [37].

As expected, the PSNR increases as the PLR decreases. However, the most important result we have obtained is that the proposed allocation scheme is able to provide the highest PSNR in all operative conditions. Notably, the proposed approach is able to guarantee a PSNR gain up to 30 dB with respect to both *LOG rule* and *EXP rule* in scenarios having more than 10 UEs.

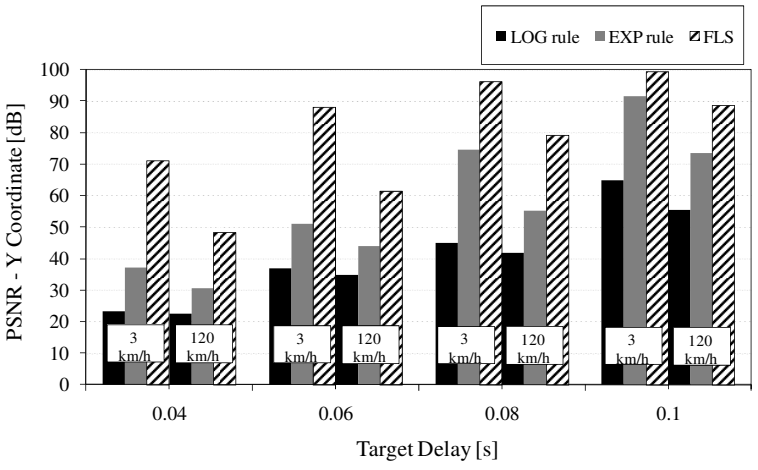
To demonstrate the effectiveness of the proposed approach, on our lab web site [5] it is possible to appreciate the quality of received video in a scenario with inter-cell interference, where users travel at 120 km/h and the target delay for real time flows has been set to 40 ms (some significant pictures taken from these videos are reported in Fig. 11). It is worth to note that, despite the critical considered scenario, the

[4]Note that in the case of zero losses a maximum PSNR value is reached, clipped to 100 dB (as suggested by the JM H264/AVC reference software encoder, http://iphome.hhi.de/suehring/tml/).

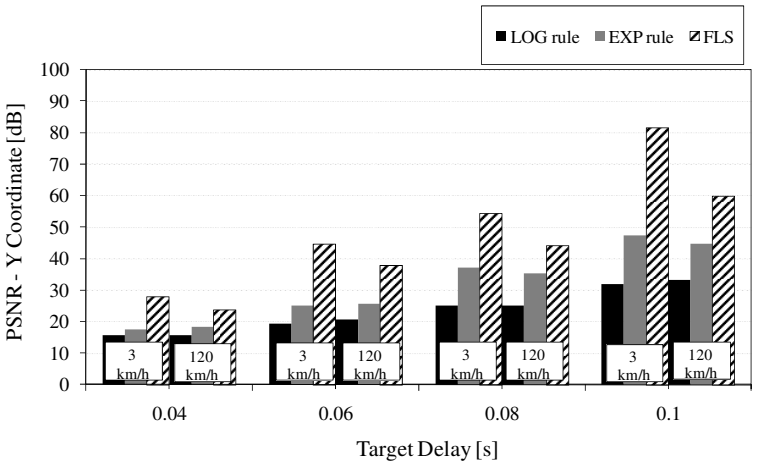
[5]http://telematics.poliba.it/QoE-on-LTE/.



(a)



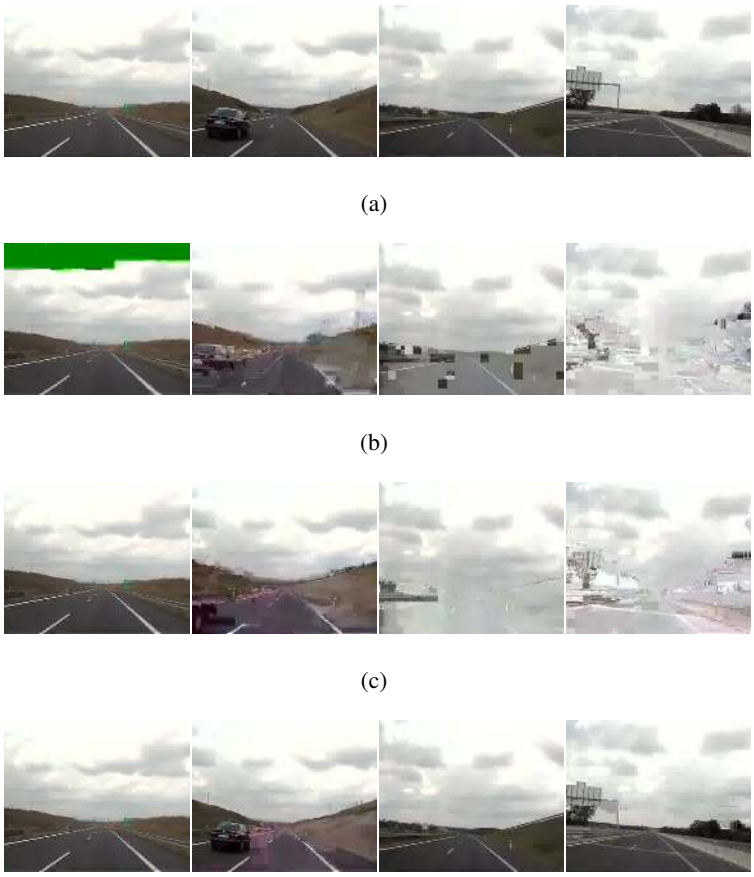
(b)



(c)

Fig. 10. PSNR of the Y component of video flows in a scenario with (a) 10, (b) 15, and (c) 20 UEs.

FLS allocation scheme is able to provide a perceived video quality better then other considered allocation schemes.



(a)

(b)

(c)

(d)

Fig. 11. Significant pictures taken from received videos (20UEs, 120 km/h, target delay = 0.04s). Sent video (a) vs received video using LOG rule (b), EXP rule (c) and FLS (d) algorithms.

VI. Conclusions

This work has considered the problem of packet scheduling for multimedia real-time flows in the downlink of LTE mobile networks. A two-level algorithm has been designed by exploiting discrete time feedback control theory. The properties of the proposed approach have been theoretically investigated to demonstrate that it is suitable to provide both real-time and best effort services. Finally, numerical simulations have been presented to confirm the analytical results. The effectiveness of the proposed approach have been highlighted comparing it with other well-known scheduling strategies and considering

both the effect of the inter-cell interference and the impact of the packet loss ratio on the QoE of real-time flows perceived by end users. Future research will consider also the more challenging problem of scheduling, at the same time, both the uplink and the downlink directions using also non-linear controllers.

APPENDIX A

Herein, we will prove Theorem 1. First of all, if eq. (4) holds, it is possible to show that the system is BIBO (Bounded Input Bounded Output) stable. In fact, we have [24]:

$$\sum_{k=0}^{+\infty} |h_{s_i}(k)| = \sum_{k=0}^{M_i} |c_i(k)| < +\infty. \tag{12}$$

Now, we can demonstrate the constraints on coefficients $c_i(n)$. To this aim, it is important to illustrate the system behavior in response to a pulse $d_i(k) = \delta(k)$. Signals $d_i(k)$, $u_i(k)$, and $q_i(k)$ are shown in Fig. 12.

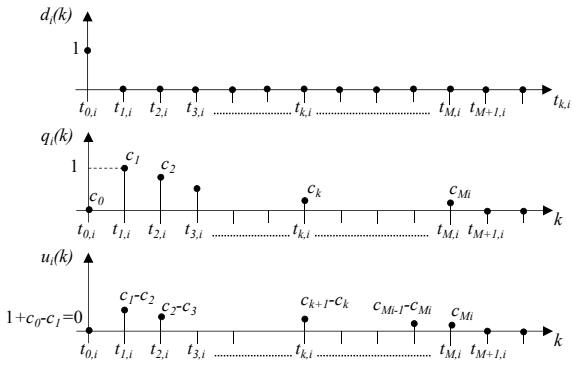


Fig. 12. FLS response to a pulse of data.

If we consider a Kronecker pulse as input to the i-th queue (this models a burst of data), obviously the queue response given by eq. (4) cannot be negative, because it represents data to transmit. Therefore, it holds that $h_{s_i}(k) \geq 0 \Leftrightarrow c_i(n) \geq 0$. Moreover, the queue cannot contain more data than its input (i.e., a pulse with width equal to 1). It means that $h_{s_i}(k) \leq 1 \Leftrightarrow c_i(n) \leq 1$.

Furthermore, to guarantee the system causality, we have to set $c_i(0) = 0$ and $c_i(1) = 1$. In fact, a pulse of data arriving during the first sampling interval $[t_{0,i}, t_{1,i}]$ will be enqueued during that interval and it will be transmitted not before the second sampling interval $[t_{1,i}, t_{2,i}]$. In other words, assuming at time $t = 0$ an empty queue, i.e., $q_i(0) = 0$, and a single data pulse as system input, i.e., $d_i(k) = \delta(k)$, we have to impose that $q_i(1) = 1$. This means, equivalently, that it should be $c_i(0) = 0$ and $c_i(1) = 1$ in eq. (4).

Now, considering the Kronecker pulse as system input, i.e., $d_i(k) = \delta(k)$, from eqs. (1) and (4) it turns out that:

$$u_i(k) = \delta(k) + \sum_{n=0}^{M_i} c_i(n)\delta(k-n) - \sum_{n=1}^{M_i} c_i(n)\delta(k+1-n). \tag{13}$$

After a bit of algebra, we obtain:

$$u_i(k) = c_i(M_i)\delta(k-M_i) + \sum_{n=1}^{M_i-1} [c_i(n) - c_i(n+1)]\delta(k-n). \tag{14}$$

Considering that $u_i(k)$ cannot be negative, it holds that $c_i(n) \geq c_i(n+1)$ for $n \geq 1$.

To summarize, in eq. (4) we have to impose the constraints

$$0 \leq c_i(n) \leq 1 \ \forall n; \qquad c_i(n) \geq c_i(n+1), n \geq 1. \tag{15}$$

Finally, we can prove that if eq. (4) is the system pulse response, it is possible to obtain bounded packet delays, smaller than $M_i + 1$ sampling intervals.

This requires that the queue backlog measured in $t_{k+1,i}$ will be transmitted in at most M_i+1 sampling interval. In this way, a generic packet that entered the queue during the time interval $[t_{k,i}, t_{k+1,i}]$ will wait in queue for at most $M_i + 1$ sampling intervals. This can be expressed as:

$$\sum_{n=0}^{M_i} u_i(k+n) \geq q_i(k) \qquad \forall k \geq 0 \tag{16}$$

which, by considering eq. (1), can be equivalently rewritten as:

$$\sum_{n=0}^{M_i} d_i(k+n) \geq q_i(k+M_i+1) \qquad \forall k \geq 0 \tag{17}$$

Considering eq. (3), we obtain:

$$q_i(k) = h_{s_i}(k) * d_i(k) = \sum_{n=0}^{M_i} c_i(n) d_i(k-n) \ . \tag{18}$$

Substituting eq. (18) in (17), the eq. (17) is equivalent to the following inequality:

$$\sum_{n=0}^{M_i} d_i(k+n) \geq \sum_{n=0}^{M_i} c_i(n) d_i(k+M_i-n+1) \tag{19}$$

Imposing $m = M_i - n + 1$, it becomes:

$$d_i(k) + \sum_{n=1}^{M_i} d_i(k+n) \geq \sum_{m=1}^{M_i} c_i(M_i - m + 1) d_i(k+m) \tag{20}$$

that is

$$d_i(k) + \sum_{n=1}^{M_i} [1 - c_i(M_i - n + 1)] d_i(k+n) \geq 0 \tag{21}$$

Remembering that $d_i(k) \geq 0$ and $0 \leq c_i(n) \leq 1$, the last inequality (21) holds for all k values. This proves the thesis.

APPENDIX B

In this appendix, we demonstrate that the eq. (4) for the system pulse response is satisfied when the transfer function of the controller is given by eq. (7).

In fact, by definition, the system transfer function $H_{S_i}(z)$ is just the $\mathcal{Z}$-transform of the system pulse response $h_{S_i}(k)$, assuming $q_i(0) = 0$. With reference to Fig. 2, we have:

$$H_{S_i}(z) = \frac{Q_i(z)}{D_i(z)} = \frac{1}{z - 1 + H_i(z)} = \mathcal{Z}[h_{S_i}(k)]. \tag{22}$$

that is, considering eq. (4):

$$\mathcal{Z}\{h_{S_i}(k)\} = \mathcal{Z}\left\{\sum_{n=0}^{M_i} c_i(n)\delta(k-n)\right\} = \sum_{n=0}^{M_i} c_i(n) z^{-n}. \tag{23}$$

REFERENCES

[1] D. McQueen, "The momentum behind LTE adoption," *IEEE Commun. Mag.*, vol. 47, no. 2, pp. 44–45, Feb. 2009.

[2] S. Khirman and P. Henriksen, "Relationship between Quality-of-Service and Quality-of-Experience for public internet service," in *Proc. of Passive and Active Measurement, PAM*, Fort Collins, CO, USA, Mar. 2002.

[3] S. Na and S. Yoo, "Allowable Propagation Delay for VoIP Calls of Acceptable Quality," in *Proc. of the First International Workshop on Advanced Internet Services and Applications*. London, UK: Springer-Verlag, 2002, pp. 47–56.

[4] G.-M. Su, Z. Han, M. Wu, and K. Liu, "Joint uplink and downlink optimization for real-time multiuser video streaming over WLANs," *IEEE J. Sel. Topics Signal Process*, vol. 1, pp. 280–294, Aug. 2007.

[5] H. Ekstrom, "QoS control in the 3GPP evolved packet system," *IEEE Commun. Mag.*, vol. 47, pp. 76–83, Feb. 2009.

[6] E. Dahlman, S. Parkvall, J. Skold, and P. Beming, *3G Evolution HSPA and LTE for Mobile Broadband*. Academic Press, 2008.

[7] G. Piro, L. A. Grieco, G. Boggia, F. Capozzi, and P. Camarda, "Simulating LTE cellular systems: an open source framework," *IEEE Trans. Veh. Technol.*, 2010, to be published.

[8] B. Sadiq, S. J. Baek, and G. de Veciana, "Delay-Optimal Opportunistic Scheduling and Approximations: The Log Rule," *IEEE/ACM Trans. on Netw.*, vol. PP, no. 99, pp. 1–14, May 2010.

[9] 3GPP, *Tech. Specif. Group Radio Access Network; Physical layer aspect for evolved Universal Terrestrial Radio Access (UTRA) (Release 7), 3GPP TS 25.814*.

[10] ——, *Tech. Specif. Group Radio Access Network; Physical Channel and Modulation (Release 8), 3GPP TS 36.211*.

[11] T. Halonen, J. Romero, and J. Melero, *GSM, GPRS and EDGE Performance: Evolution Towards 3G/UMTS*. Wiley, 2003.

[12] L. Nuaymi, *WiMAX: Technology for Broadband Wireless Access*. Wiley, 2008.

[13] P. Kela, J. Puttonen, N. Kolehmainen, T. Ristaniemi, T. Henttonen, and M. Moisio, "Dynamic packet scheduling performance in UTRA Long Term Evolution downlink," in *Proc. of Int. Symposium on Wireless Pervasive Computing, ISWPC*, May 2008.

[14] 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 *Proc. of IEEE Veh. Tech. Conf., VTC-Spring*, Marina Bay, Singapore, May 2008.

[15] Y. Lin and G. Yue, "Channel-adapted and buffer-aware packet scheduling in LTE wireless communication system," in *Proc. of Int. Conf. on Wireless Communications, Networking and Mobile Computing, WiCOM*, Dalian China, Oct. 2008.

[16] R. Kwan, C. Leung, and J. Zhang, "Proportional fair multiuser scheduling in LTE," *Proc. of IEEE Signal Processing Letters*, vol. 16, no. 6, pp. 461–464, Jun. 2009.

[17] H. Luo, S. Ci, D. Wu, J. Wu, and H. Tang, "Quality-driven cross-layer optimized video delivery over LTE," *Communications Magazine, IEEE*, vol. 48, no. 2, pp. 102 –109, feb. 2010.

[18] F. A. Bokhari, H. Yanikomeroglu, W. K. Wong, and M. Rahman, "Cross-layer resource scheduling for video traffic in the downlink of OFDMA-based wireless 4G networks," *EURASIP J. Wirel. Commun. Netw.*, pp. 1–10, 2009.

[19] J. Park, S. Hwang, and H. S. Cho, "A packet scheduling scheme to support real-time traffic in OFDMA systems," in *Proc. of IEEE Veh. Tech. Conf., VTC-Spring*, Dublin Ireland, Apr. 2007.

[20] M. Assaad, "Frequency-Time Scheduling for streaming services in OFDMA systems," in *Wireless Days, 2008. WD '08. 1st IFIP*, 2008, pp. 1 –5.

[21] Y. Qian, C. Ren, S. Tang, and M. Chen, "Multi-service QoS guaranteed based downlink cross-layer resource block allocation algorithm in LTE systems," in *Wireless Communications Signal Processing, 2009. WCSP 2009. International Conference on*, 2009, pp. 1 –4.

[22] H. Ramli, R. Basukala, K. Sandrasegaran, and R. Patachaianand, "Performance of well known packet scheduling algorithms in the downlink 3GPP LTE system," in *Communications (MICC), 2009 IEEE 9th Malaysia International Conference on*, 2009, pp. 815 –820.

[23] X. Wang, G. B. Giannakis, and A. G. Marques, "A Unified Approach to QoS-Guaranteed Scheduling for Channel-Adaptive Wireless Networks," *Proceedings of the IEEE*, vol. 95, no. 12, pp. 2410 – 2431, Dec. 2007.

[24] K. J. Astrom and B. Wittenmark, *Computer controlled systems: theory and design*, 3rd ed. Prentice Hall, 1995.

[25] B. Sadiq, R. Madan, and A. Sampath, "Downlink scheduling for multiclass traffic in LTE," *EURASIP J. Wirel. Commun. Netw.*, vol. 2009, pp. 9–9, 2009.

[26] 3GPP, *Tech. Specif. Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) radio transmission and reception (Release 9), 3GPP TS 36.101*.

[27] T. Camp, J. Boleng, and V. Davies, "A survey of mobility models for ad hoc network research," *Wireless Commun. and Mobile Computing*, vol. 2, pp. 483–502, 2002.

[28] G. Boggia, P. Camarda, L. A. Grieco, and S. Mascolo, "Feedback-based control for providing real-time services with the 802.11e mac," *IEEE/ACM Trans. on Netw.*, vol. 15, no. 2, pp. 323–333, Apr. 2007.

[29] 3GPP, *Tech. Specif. Group Radio Access Network; Conveying MCS and TB size via PDCCH, 3GPP TSG-RAN WG1 R1-081483*.

[30] Y. R. Zheng and C. Xiao, "Simulation models with correct statistical properties for rayleigh fading channels," *IEEE Trans. on Comm.*, vol. 2, no. 6, pp. 920 – 928, Jun. 2003.

[31] S. Sesia, M. P. J. Baker, and I. Toufik, *LTE, the UMTS long term evolution: from theory to practice*. John Wiley and Sons, 2009.

[32] 3GPP, *Tech. Specif. Group Services and System Aspects; Policy and charging control architecture (Release 11), 3GPP TS 23.203*.

[33] S. Shakkottai and A. Stolyar, "Scheduling algorithms for a mixture of real-time and non-real-time data in hdr," in *17th International Teletraffic Congress (ITC '01)*, 2001.

[34] M. Mardiak and J. Polec, “Novel method for objectively measuring video quality,” in *Proc. of the 53th Intenrational Symposium ELMAR*, Sep. 2010.

[35] G. Boggia, P. Camarda, R. Dell’Aquila, O. Fiume, and L. A. Grieco, “CF-MAC Protocol for Voice Communication in Wireless Ad-Hoc Networks,” in *Proc. of Second Workshop on multiMedia Applications over Wireless Networks, MediaWiN 2007*, Aveiro, Portugal, Jul 2007.

[36] International Telecommunication Union (ITU), *Definition of categories of speech transmission quality*, ITU-T Recommendation G.109, Aug. 1996.

[37] Z. Zhe and H. Wu, “A Ratio Sensitive Quality Metric for Digital Video,” in *Proc of the First International Workshop on Video Processing and Quality Metrics for Consumer Electronics, VPQM*, Scottsdale, Arizona, USA, Jan. 2005.