A Two-level Scheduling Algorithm for QoS Support in the Downlink of LTE cellular networks

Giuseppe Piro, Luigi Alfredo Grieco, Gennaro Boggia, and Pietro Camarda
DEE - Dipartimento di Elettrotecnica ed Elettronica
Politecnico di Bari
v. Orabona 4 - 70125, Bari, Italy
Email: {g.piro, a.grieco, g.boggia, camarda}@poliba.it

Abstract

Long Term Evolution represents an emerging and promising technology for providing a broadband ubiquitous Internet access. But several aspects have to be considered in order to provide an effective service to users. In particular, in this work, we consider the problem of optimizing the performance of real time downlink communications using a novel two-level scheduling algorithm. The upper level exploits an innovative approach based on discrete-time linear control theory. At the lower level, instead, a maximum throughput scheduler has been properly tailored to our purposes. The performance and the complexity of the proposed scheme have been evaluated theoretically and by using simulations. Both the analyses demonstrate the effectiveness of the proposed approach.

I. Introduction

To face the ever growing demand for packet-based mobile broadband systems, the 3GPP [1] has introduced the LTE (Long Term Evolution) specifications [2] as the next step of the current 3G mobile networks. An enhanced access network (i.e., the E-UTRAN, Evolved-UMTS Terrestrial Radio Access Network) and an evolved core network have been defined [3]. 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 worlds 3.5 billion mobile subscribers) and more than 32 million LTE subscribers are forecast by 2013 [4]. 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.

In this work, we consider the problem of optimizing the scheduling of downlink communications. In general, the most important objective of LTE scheduling is to satisfy the Quality of Service (QoS) requirements of all user by trying to reach, at the same time, an optimal tradeoff between utilization and fairness [5]. This goal is very challenging, especially in the presence of real time multimedia applications, which are characterized by strict constraints on packet delay and jitter.

To provide QoS differentiation among User Equipments (UEs) and among different traffic flows on the same UE, a bearer is introduced as an identifier for each flow requiring a particular policy [6]. The scheduler classifies packets belonging to a given bearer (i.e., a specified flow with QoS requirements) using a packet filter based on the classical five-tuple: source and destination IPs, source and destination ports, protocol identifier. Thus, packet schedulers can allocate radio resources on a per-flow base. To each bearer is associated a QoS Class Identifier, that describe the QoS level expected from the considered data flow, and up to four parameters: service class, priority, target delay, and packet loss ratio. Two kinds of bearers are defined [7]: *guaranteed bit rate* and *non-guaranteed bit rate*. The former is used to map *non real time* flows (e.g., e-mail, http web browsing, which require maximum allowable bit error rate and a minimum data rate, but no delay bounds) and *real time* flows (e.g., VoIP calls, video conferencing and streaming, which require guarantees on bit error rate, throughput, and delay). The latter is used to transfer application signaling and *best effort* traffic.

It is important to remark that in LTE networks uplink and downlink behaves differently. As a consequence, different scheduling strategies have been proposed for them by the scientific community (see Sec. III for a summary of related works). Anyway, 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 downlink scheduling strategy, that can provide strict delay bounds to real time flows. The approach exploits a two level scheme. Following LTE specifications, in our approach the time is seen as an endless sequence of frames, which are further splitted in many Transmission Time Intervals (TTIs).

At the highest level, an innovative 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, in order to satisfy its delay constraint. Then, the lowest level scheduler assigns radio resources according to the known Maximum Throughput algorithm [5] subject to the constraint imposed by FLS. Radio resources left free by real time flows can be used to provide a best effort service by the well known Proportional Fair (PF) algorithm [5], to provide a fair best effort service.

The performance of the proposed resource allocation scheme has been verified using a LTE system level simulator, written in the Matlab environment [8]. Simulation results clearly show that the developed algorithm meets the sharp delay requirements of real time users, demonstrating that the same goal cannot be met using the simple PF scheduler.

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

II. LTE Overview

The requirements of LTE networks [5] are very ambitious: 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 the packet scheduling, the link adaption, the Hybrid ARQ of packets.

At the physical layer, as many existing wireless broadband systems, the LTE radio interface supports several duplexing techniques: the frequency division duplex, the time division duplex, and the half frequency division duplex.

The radio transmissions in LTE 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 users. Moreover, 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 fading of the radio channel. On the other hand, SC-FDMA is used in the LTE uplink to increase the power efficiency of 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].

Each LTE frame lasts $T_f = 10$ ms and it is divided into equally size sub-frame, called Transmission Time Interval (TTI), lasting 1 ms. The whole bandwidth is divided into 180 kHz physical Resource Blocks (RBs), each one lasting 0.5 ms and consisting of 6 or 7 symbols in the time domain (according to the OFDM prefix-code duration) and 12 consecutive sub-carriers in the frequency domain [10]. The resources allocation is realized every TTI, that is exactly every two consecutive resource blocks; thus, resource allocation is done on a resource block pair basis.

The packet scheduler and link adaptation modules work together and run both in the base station (the so called evolved node B, eNodeB). Every TTI, the UE measures the channel quality for each RB in terms of Signal to Interference and Noise Ratio (SINR) and send it to the eNodeB.

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 suited modulation scheme and coding rate at the physical level, in order to maximize the spectral efficiency. This approach is known as *Adaptive Modulation and Coding* 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 can establish the maximum available physical data rate for each UE based on the received channel quality information, providing 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 throttling the amount of radio resources assigned to each user. Anyway, the issue of scheduling resource blocks in order to meet the expected QoS is very challenging in the LTE system and more in general in wireless networks. In fact, the problem of finding a simple algorithm that assigns time slots and frequency carriers to users by taking into account the expected QoS level, the behavior of data sources, and the channel status, has attracted the attention of many researchers of the field (see the survey [7] for a comprehensive view on this subject). The problem becomes more complex in the presence of users with different requirements in term of bandwidth, tolerance to delay, and reliability.

Uplink and downlink behave differently in LTE networks. As a consequence, they adopt different scheduling strategies. In the present review of papers in literature, we will describe those designed for downlink, because more strictly related to our work. In the downlink there are no constraints on the contiguity of the radio resource blocks. Therefore, different algorithms have been proposed.

In [13]–[16] scheduling strategies based on PF algorithm are proposed to target high network throughput and fairness between users. Despite their efficiency, they do not take into account the strict delay requirements of real-time flows.

In [17], an efficient scheduling strategy for VoIP flows has been proposed. Its key idea consists in the dynamic activation of a VoIP priority mode and the adaptation of its duration in order to minimize the performance degradation of the overall system. No mention to other kinds of real-time flows has been done in the paper.

Scheduling algorithms proposed in [18]–[21] are focused on real-time services and their contributions are very relevant to our discussion.

In [18], a downlink scheduling algorithm for OFDMA systems is proposed, which exploits a mathematical formulation based on utility functions to provide guarantees on packet drop ratio and play-out outage ratio in video streaming services.

A packet scheduling scheme that supports real time traffic with multi-level delay constraints is proposed in [19] . Moreover, under the delay constraint, a throughput enhancement strategy is jointly conceived. The effectiveness of the scheme has been demonstrated in comparison with the Exponential Scheduling scheme, the Modified Largest Weighted Delay First scheme, and the Round robin scheme. In each TTI, the algorithm privileges flows whose deadline is expiring, thus pushing packet transmissions very close to their deadlines. Since some flows with pending data near to the deadline could experience a sudden decrease of the channel quality, a violation of the target delay could arise.

In [20], a generic scheduling for channel-adaptive wireless networks is proposed to provide absolute delay guarantees to real-time flows. This scheme presents a very high computational complexity and is hard to adapt to LTE radio interface.

Finally, in [21] a cross-layer resource allocation scheme for cellular networks is proposed. It is composed by two modules: the scheduler and the allocator. The former selects flows to schedule and defines the bandwidth to allocate to each of them. The latter determines which sub-channel should be assigned to each flow. Despite the allocator module has been designed to exploit frequency diversity of the OFDMA radio interface, no sophisticated scheduling algorithms have been proposed to efficiently serve real-time flows.

IV. Two-level scheduling

This section describes how our two-level scheduling algorithm for the LTE downlink has been designed. This composite algorithm is made of two distinct mechanisms, which properly interact together, to support both real time and best effort services (see Fig. 1).

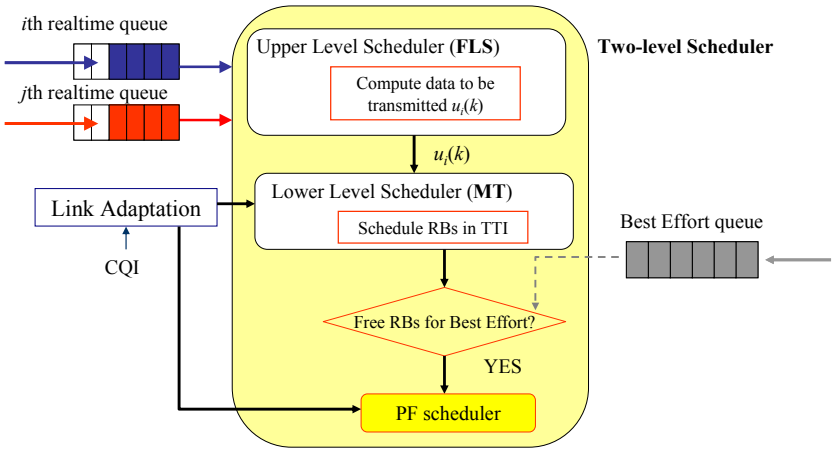

Fig. 1. The two-level scheduling algorithm.

The proposed approach is able to provide strict delay bounds to real time flows. 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. It has been designed assuming a very common network architecture made by: a shared channel, N transmission queues, and a scheduling algorithm that, knowing transmission queue levels, properly distributes the channel bandwidth keeping strict bounds on queuing delays. To achieve a small computational complexity, FLS has been designed using the linear discrete-time control theory [22].

Once FLS has accomplished this task, the lowest layer scheduler, every TTI, assigns RBs according to the Maximum Throughput algorithm according to the constraints 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 source. The lowest layer scheduler, instead, allocates resource blocks in each TTI in order to maximize the 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 sources related to UEs experiencing the best channel quality and then (i.e., when these sources have transmitted the amount of data imposed by FLS) it considers the remaining ones.

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. It is important to underline that the TTIs (and then the RBs) in which each real time source should actually transmit its packets is decided by the lowest level of our allocation scheme.

A. The upper level of the scheduler

The FLS scheduler (that is, the upper level of or two-level scheduler) has been designed using discrete-time linear control theory elements. In our system, we suppose that N active traffic flows share the wireless channel. Associated to each of these flows, there is a queue, where packets are stored waiting for transmission. The FLS scheduler evaluates the transmission needs of each queue every LTE frame.

We define the starting time, $t_{k,i}$, of the k-th frame; it is considered as the constant sampling instant in our system, i.e., $\Delta t(k) = t_{k+1,i} - t_{k,i}$ is the sampling interval. More precisely, in our system, the sampling time $\Delta t(k)$ is equal to 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.

1) The control law: The FLS is an original scheduler based on a control law that has to compute, at the beginning of the k-th frame, the quota of data $u_i(k)$ that the i-th flow at a given UE should transmits in that frame. Such a control law should be properly designed to provide bounded delays to transmitted packets, assuring, at the same time, BIBO stability [22] to the system defined by eq. (1).

We will assume the following general control law:

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

where the '$*$' operator is the discrete time convolution [22].

Eq. (2) means that the amount of data to be transmitted 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 transfer function $H_i(z) = \mathcal{Z}[h_i(k)]$ [22].

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 $H_i(z)$. In the following, the pulse response of the system will be referred to as $h_{s_i}(k)$, so that the following equality holds:

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

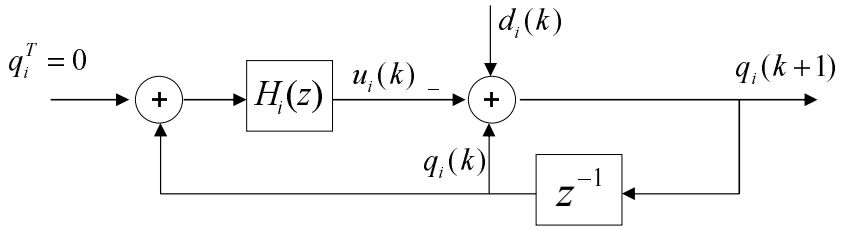

Fig. 2. Control loop of the allocation algorithm at the upper layer of the scheduler.

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 stability to the system and guaranteed queuing delays. These constraints could be fulfilled if the closed-loop response to the Kronecker pulse $\delta(k)$ [22] (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 $c_i(n)$ are real finite coefficients, i.e., $c_i(n) \in R$, and M_i is the length of the pulse response.

In fact, if eq. (4) holds, the system is BIBO stable because [22]:

$$\sum_{k=0}^{+\infty} |h_{s_i}(k)| = \sum_{k=0}^{M_i} |c_i(k)| < +\infty. \tag{5}$$

To clearly explain the system behavior in response to a pulse $d_i(k) = \delta(k)$, the signals $d_i(k)$, $u_i(k)$, and $q_i(k)$ are shown in Fig. 3.

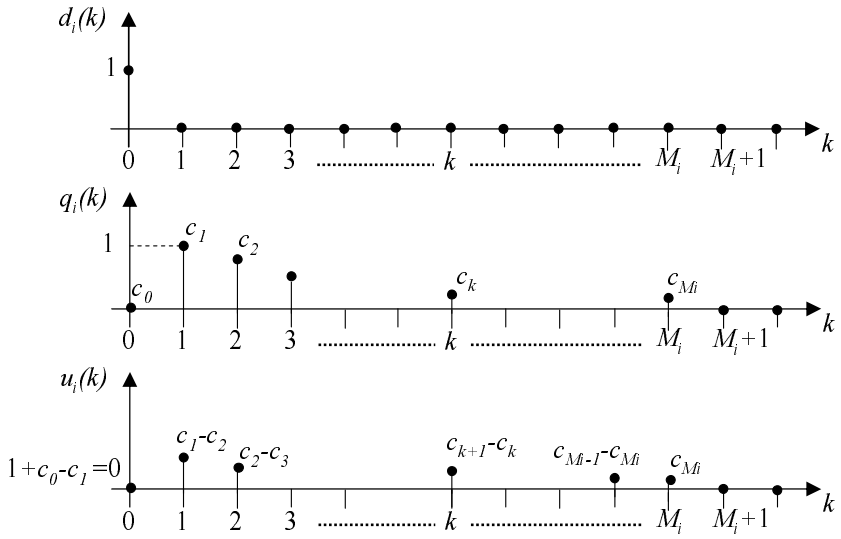

Fig. 3. FLS response to a pulse of data.

If we consider a Kronecker pulse as input to the queue, obviously the queue response given by eq. (4) cannot be negative. 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 $h_{s_i}(k) \leq 1 \Leftrightarrow c_i(n) \leq 1$.

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. 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 the queue is filled only by the data pulse at time $t = 1$, i.e., $q_i(1) = 1$. 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:

$$u_i(k) = \delta(k) + \sum_{n=0}^{M_i} c_i(n)\delta(k-n) - \sum_{n=1}^{M_i} c_i(n)\delta(k+1-n). \tag{6}$$

After a bit of algebra, we obtain:

$$u_i(k) = c_i(M_i)\delta(k-M_i) + \sum_{n=1}^{M_i-1}[c_i(n) - c_i(n+1)]\delta(k-n). \tag{7}$$

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

$$0 \leq c_i(n) \leq 1 \ \forall n; \qquad c_i(n) \geq c_i(n+1), n \geq 1. \tag{8}$$

It will be mathematically demonstrated later that these constraints are able to provide upper bounded queuing delays.

Proposition 1: The eq. (4) for the system pulse response is satisfied when the transfer function of the controller is:

$$H_i(z) = \frac{U_i(z)}{Q_i(z)} = \left[(1-z)\sum_{n=0}^{M_i} c_i(n)z^{-n} + 1\right] \Big/ \sum_{n=0}^{M_i} c_i(n)z^{-n}. \tag{9}$$

Proof: 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:

$$H_{S_i}(z) = \frac{Q_i(z)}{D_i(z)} = \frac{1}{z - 1 + H_i(z)} = \mathcal{Z}[h_{S_i}(k)]. \tag{10}$$

that is, considering eq. (4):

$$\mathcal{Z}\{h_{S_i}(k)\} = \mathcal{Z}\left\{\sum_{n=0}^{M_i} c_i(n)\delta(k-n)\right\} = \sum_{n=0}^{M_i} c_i(n)z^{-n}. \tag{11}$$

Solving eq. (10) with respect to $H_i(z)$ the proof is obtained.

Theorem 1: The queuing delay τ_i of the i-th queue is smaller than M_i+1 sampling intervals. Considering that each sampling interval lasts T_f, the upper bound of the delays is:

$$\tau_i = (M_i + 1)T_f \ . \tag{12}$$

Proof: The thesis 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:

$$\sum_{n=0}^{M_i} u_i(k+n) \geq q_i(k) \qquad \forall k \geq 0 \tag{13}$$

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

$$\sum_{n=0}^{M_i} d_i(k+n) \geq q_i(k+M_i+1) \qquad \forall k \geq 0 \tag{14}$$

Transforming back to time domain eq. (10), we obtain:

$$q_i(k) = \sum_{n=0}^{M_i} c_i(n) d_i(k-n) \ . \tag{15}$$

Substituting eq. (15) in (14), the eq. (14) is equivalent to the following inequality:

$$\sum_{n=0}^{M_i} d_i(k+n) \geq \sum_{n=0}^{M_i} c_i(n) d_i(k+M_i-n+1) \tag{16}$$

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

$$d_i(k) + \sum_{n=1}^{M_i} d_i(k+n) \geq \sum_{m=1}^{M_i} c_i(M_i-m+1) d_i(k+m) \tag{17}$$

that is

$$d_i(k) + \sum_{n=1}^{M_i} [1 - c_i(M_i-n+1)] d_i(k+n) \geq 0 \tag{18}$$

Remembering that $d_i(k) \geq 0$ and $0 \leq c_i(n) \leq 1$, the last inequality (18) holds for all k values. This proves the thesis.

It is important to note that it is very simple to implement the FLS algorithm in the LTE downlink scheduler. In fact, each eNodeB knows always the transmission queue of each active flows in downlink. So that, it can easily run the FLS control law every T_f. The simplicity and the mathematical validation of FLS scheduler exalt the importance of the proposed allocation scheme. In Sec. V, simulation results will confirm the effectiveness of our proposed approach.

Regarding the computational complexity of FLS, algorithm, for each active real time flow in the downlink, the eNodeB computes the amount of data $u_i(k)$ by the following control law:

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

that can be obtained by considering eq. (9) in the time domain.

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 $i = 1, \ldots, N$.

B. The lower level of the scheduler

The TTIs (and then the RBs) in which each real time source should actually transmit its packets is decided by the lowest level of our allocation scheme. For each one of the ten TTIs forming a given frame, the lower level scheduler allocates the RBs to real time flows. It considers only flows that have not yet transmitted their quota $u_i(k)$ in the previous TTIs of the same frame. Thus, when a real time source has transmitted in a frame at least the whole quota $u_i(k)$ defined by FLS scheduler, it loses the opportunity to transmit until the beginning of next frame, also if it is related to the UE with the best channel quality.

To obtain the highest spectral efficiency, at the lower level we exploits the MT algorithm [5]. The role of such an algorithm is to assign one or more RBs to those UEs that measure the highest channel quality. In our LTE system, we have considered that the CQI is measured every TTI and that the MT algorithm assigns RBs considering UEs channel quality measured in the

previous TTI. In particular, the MT algorithm in each TTI assigns each RB to that UE that presents the maximum instantaneous supportable data rate in that RB, computed by the link adaptation module using feedbacks on channel quality (see Sec. I).

It is important to note that fixing the quota $u_i(k)$, the number of the OFDM symbols (or equivalently the number of RBs) required to transmit the data in the frame 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 can be used to provide the best effort service. In our study we have scheduled the resources available for best effort using PF scheduler [5]. It is important to remark that MT algorithm is not able to provide the fairness between best effort flows. Instead, the PF has been designed both to maximize total network throughput and to guarantee the fairness between BE services. For this reason, we adopt PF for the best effort flows.

The PF algorithm assigns the shared radio resources to the users that present the relatively best radio channel condition. To be more precise, in our proposed allocation scheme, PF assigns RBs to downlink connections belonging to UEs presenting the best ratio, w, of instantaneous available data rate to average data rate. That is, with reference to the i-th UE in the j-th sub-band:

$$w_{i,j} = R^M_{i,j}/\overline{R}_{i,j}, \tag{19}$$

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

Fig. 4 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 lost their rights to transmit during the current frame.

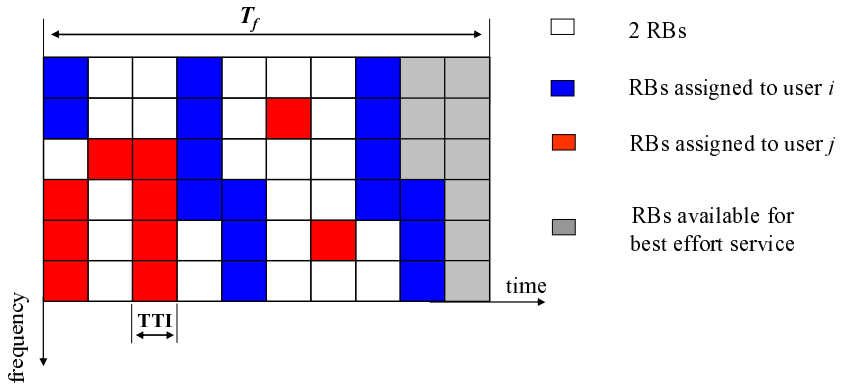

Fig. 4. Example of the resource allocation scheme.

It is important to note that real time flows are scheduled independently from best effort services; as a consequence, one could adopt the scheduler it prefers to provide best effort service without affecting real time applications.

V. Performance Evaluation

To study the effectiveness of the proposed downlink resource allocation scheme, a LTE MAC system level simulation has been developed with an ad hoc simulator written in the Matlab environment. Packet scheduling analysis is based on E-UTRAN LTE downlink parameters proposed by the 3GPP specification [9].

Simulation results will demonstrate the respect of target delays, equal to $(M_i + 1)T_f$ (see Teorem 1), in all operative conditions, exalting the ability of proposed resource allocation scheme to provide sharp delay bounds to real time flows.

The simulation system scenario consists of one eNodeB and a variable number of UEs (see Fig. 5). The traffic transmitted by the eNodeB on the downlink is made by a mix of 60% of voice flows encoded with the G.729 [23] standard, 20% of MPEG-4 [24] encoded video flows, and 20% of data flows related to best effort service. Each UE hosts a single multimedia or a receiver of best effort traffic. UEs are uniformly distributed in a LTE cell.

Our simulations are based on realistic traffic models. For the video flows, we have used traffic traces available from the video trace library [25]. For G.729 voice flows, instead, 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 [26]. 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 source. The main characteristics of the considered multimedia flows are summarized in Tab. I.

We have considered scenarios with 15, 20, and 25 UEs. Two type of mobility models has been considered for UEs according to [9]: 1) pedestrian with user speed equal to 3 km/h; and 2) vehicular with user speed equal to 30 km/h. Each user follows the former with probability 0.7 and the latter with probability 0.3.

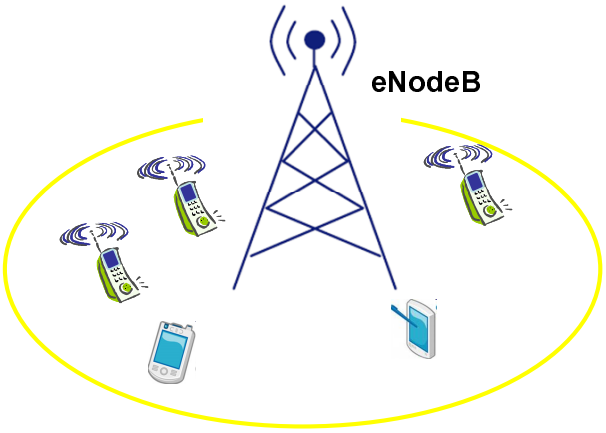

Fig. 5. Scenario with multimedia flows.

TABLE I
MAIN FEATURES OF THE CONSIDERED MULTIMEDIA FLOWS.

Features	Flow Type	
	MPEG-4	**G.729**
Nominal MSDU [byte]	1536	60
Max. MSDU [byte]	2304	60
Burst Size [byte]	16745	60
Mean Data Rate [kbps]	770	8.4
Peak Data Rate [kbps]	3300	24

For what concerns the channel quality measured by UE, in our simulations SINR has been mathematically obtained considering eNodeB power transmission, fast fading, path loss shadowing and noise, as defined in [9]. The fast fading has been generated by Jakes Model [27]. The path loss is given by $P_L = 128.1 + 37.6 \log d$, where d is the distance between eNodeB and UE in kilometers. The large scale shadowing fading has been modeled trough a log-normal distribution with 0 mean and 8 dB standard deviation.

The link adaptation uses the channel quality value to get, through the Shannon theorem [28], the maximum instantaneous supportable data rate, $R_{i,j}^M$, for $i-th$ user and $j-th$ subband:

$$R_{i,j}^M = B_j \log_2 \left(1 + \frac{SINR(i,j)}{\Gamma}\right), \tag{20}$$

where B is the bandwidth of a RB (i.e., 180 kHz) and Γ is a coefficient introduced for considering the difference between practical implementation and theoretical results of Shannon theorem and depending on the target BER [29]:

$$\Gamma = -\ln(5 \cdot BER)/1.5. \tag{21}$$

The link adaptation module chooses the most effective modulation scheme among QPSK, 16QAM and 64QAM. Knowing the bandwidth of each RB, a fixed value of spectral efficiency (or equivalently a fixed physical data rate) corresponds to every digital modulation scheme. Then, the packet scheduler assigns RBs according to resource allocation scheme rules. The main simulation parameters are reported in Tab.II.

For FLS scheduler, we have set $M_i = 3$ or $M_i = 5$. With reference to eq. (4), we set the $c_i(n)$ coefficients as follows:

$$c_i(0) = 0; \quad c_i(n) = 1 - (n-1)/M_i \quad \forall n = 1, \ldots, M_i. \tag{22}$$

In this way $H_{S_i}(z)$, described in eq. (11), has a linear pulse shaping. Accordingly, the filter $H_i(z)$ has been designed with the given coefficient $c_i(n)$ by using eq. (9).

For best effort service, algorithm have been used. It is important to remark that the behavior of FLS scheduler is independent from the scheduler used for best effort flows. Moreover, to provide a further insight, we have compared FLS with an allocation scheme that use only PF algorithm for all service flows.

Figs. 6-8 show the Cumulative Distribution Functions (CDFs) of packet delays of real time flows.

It is worth noting that our proposed resource allocation scheme is able to provide bounded delays for multimedia flows in all considered operative conditions. In details, using both $M_i = 3$ and $M_i = 5$, the target delay (equal to 40 ms and 60 ms, respectively) is always satisfied. On the other hand, delays obtained using the PF algorithm alone are very wide, especially for MPEG-4 flows. Moreover, we can observe that G.729 flows exhibit a smaller packet delay with respect to MPEG-4 ones. The reason is that voice flows have smaller packets than MPEG-4 ones.

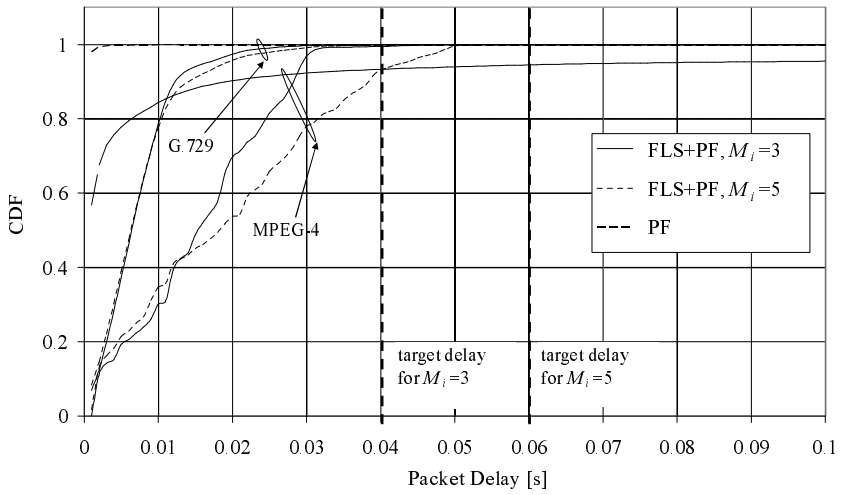

Fig. 6. CDF of packet delay in a scenario with 15 flows.

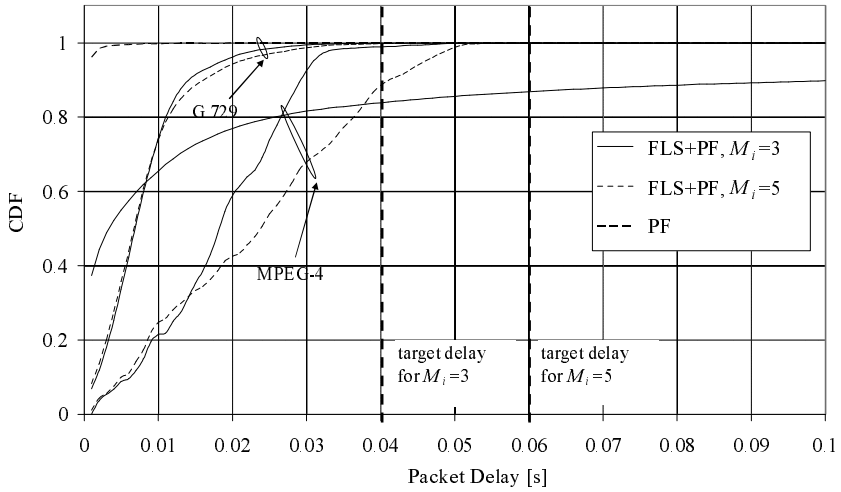

Fig. 7. CDF of packet delay in a scenario with 20 flows.

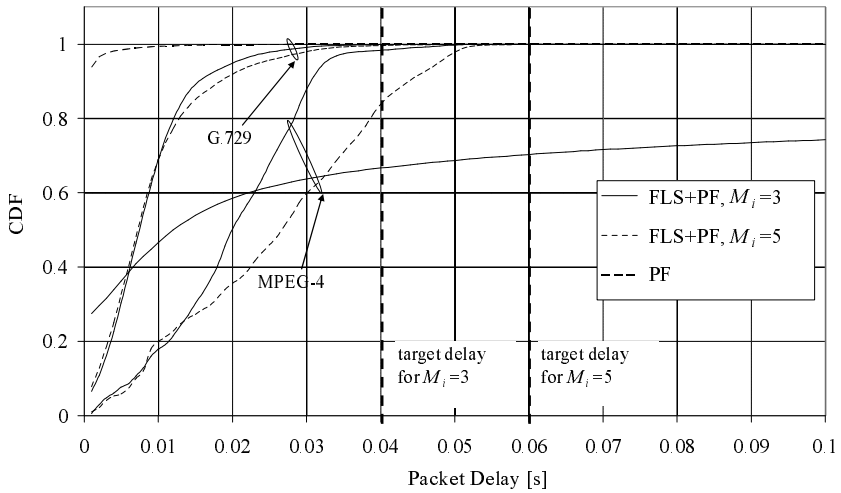

Fig. 8. CDF of packet delay in a scenario with 25 flows.

TABLE II
SIMULATION PARAMETERS.

Parameter	Value
Simulation length	180 s
Physical Detail	Symbol for TTI: 14; SubFrame length: 1 ms; SubCarries per RB: 12 SubCarrier spacing: 15 kHz; Bandwidth: 5 MHz; eNodeB: Power transmission=46 dBm; omnidirectional antenna; MIMO: off Modulation Scheme: QPSK, 16QAM and 64QAM Coding Rate for Physica Downlink Channel: 1/3 BER: 10^{-5}
Overhead	RTP/UDP/IP with ROCH compression: 3 byte MAC and RLC: 5 byte; PDCP: 2 byte; CRC: 3 byte L1/L2: 3 symbols
Cell layout	radius: 500 m
HARQ	off
CQI	Measured period: 1 ms; Number of RBs per CQI: 2
Number of UEs	15, 20, 25
Traffic Model	real time traffic type: MPEG (20%), VoIP (60%) best effort flows: infinite buffer (20%)

To study the behavior of best effort flows, fig. 9 shows the average goodput, defined as the rate of useful bits successfully transmitted by this kind of flows during the whole simulation. As expected, the goodput decreases as the number of UEs increases, because of the smaller available bandwidth left free for best effort service. Moreover, we note that, when using only the PF algorithm for all flows, we obtain a higher goodput for best effort flows with respect to the use of the FLS algorithm. The reason is that PF does not provide any guarantees to real-time flows, thus leaving more bandwidth free for best effort ones.

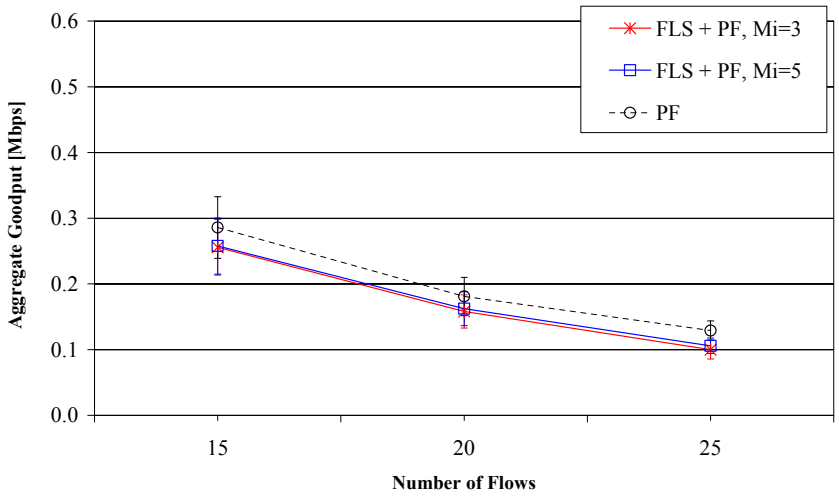

Fig. 9. Goodput achieved by best effort flows.

To evaluate the fairness of the best effort service, we have also computed the Jain Fairness Index [30], considering the average goodput achieved by these flows at the end of each simulation. In all operative conditions the index is very close to 0.9 (see Fig. 10), meaning that all considered scheduling strategies provide comparable levels of fairness.

VI. CONCLUSION

This work has considered the problem of packet scheduling 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 in order to confirm the analytical results. Future research will consider also the more challenging problem of scheduling, at the same time, both the uplink and the downlink directions. Furthermore, the comparison with respect to other advanced scheduling algorithms will be also afforded.

ACKNOWLEDGMENT

This work was funded by projects PS-121 and PS-092 (Apulia Region, Italy).

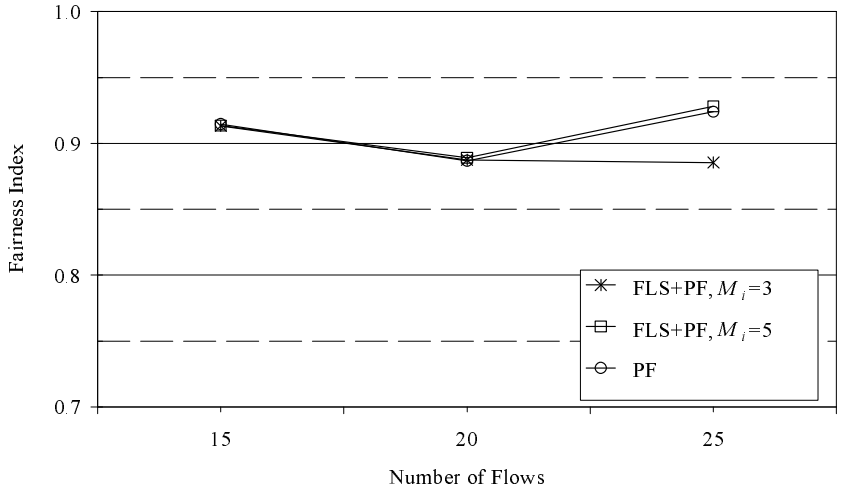

Fig. 10. Fairness Index for best effort flows.

REFERENCES

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

[2] 3GPP, *Tech. Specif. Group Radio Access Network Requirements for Evolved UTRA (E-UTRA) and Evolved UTRAN (E-UTRAN), 3GPP TS 25.913*.

[3] ——, *Tech. Specif. Group Services and System Aspects Service Requirements for Evolution of the 3GPP System (Release 8), 3GPP TS 22.278*.

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

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

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

[7] 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.

[8] Matlab, http://www.mathworks.com/.

[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] 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.

[14] 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.

[15] 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.

[16] 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.

[17] S. Choi, K. Jun, Y. Shin, S. Kang, and B. Choi, "MAC Scheduling Scheme for VoIP Traffic Service in 3G LTE," in *Proc. of IEEE Veh. Tech. Conf., VTC-Fall*, Baltimore, MD, USA, Oct. 2007.

[18] H. Lei, C. Fan, X. Zhang, and D. Yang, "QoS aware packet scheduling algorithm for OFDMA systems," in *Proc. of IEEE Veh. Tech. Conf., VTC-Fall*, Baltimore, MD, USA, Oct. 2007.

[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] 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.

[21] L. Badia, A. Baiocchi, A. Todini, G. Merlin, S. Pupolin, A. Zanella, and M. Zorzi, "On the Impact of Physical Layer Awareness on Scheduling and Resource Allocation in Broadband Multicellular IEEE 802.16 Systems," in *IEEE Wireless Communications*, vol. 14, no. 1, 2007, pp. 36 – 43.

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

[23] International Telecommunication Union (ITU), *A silence compression scheme for G.729 optimized for terminals conforming to Recommendation V.70*, ITU-T Recommendation G.729 Annex B, Nov. 1996.

[24] MPEG-4 Video Group, "Mpeg-4 overview," available at http://mpeg.telecomitalialab.com/, Mar. 2002.

[25] "Video trace library," available at http://trace.eas.asu.edu/.

[26] C. Chuah and R. H. Katz, "Characterizing Packet Audio Streams from Internet Multimedia Applications," in *Proc. of IEEE ICC*, New York, USA, Apr. 2002, pp. 1199 – 1203.

[27] C. Wei, Z. Lili, and H. Zhiyi, "Second-order Statistics of Improved Jakes' Models for Rayleigh Fading Channels Wireless Communications," in *Proc. of Int. Conf. on Networking and Mobile Computing, WiCom*, Shanghai, China, Sep. 2007, pp. 1108 – 1111.

[28] J. Proakis, *Digital Communications, IV Edition*. McGraw-Hill, 2002.

[29] H. Seo and B. G. Lee, "A proportional-fair power allocation scheme for fair and efficient multiuser ofdm systems," in *Proc. of IEEE GLOBECOM*, Dallas, TX, USA, Dec. 2004.

[30] R. Jain, *The Art of Computer Systems Performance Analysis*. John Wiley & Sons, 1991.