On finding the right balance between fairness and efficiency in WiMAX scheduling through analytical modeling

Sébastien Doirieux, Bruno Baynat, Thomas Begin
Université Pierre et Marie Curie - Paris, FRANCE, {firstname.lastname}@lip6.fr

***Abstract*— In this paper, we explore a way to find the right scheduling policy for WiMAX networks, that achieves the best compromise between an efficient use of the resource and a relative fairness among users. This problem is of primary importance as no scheduling policy has been recommended in the WiMAX standard. To do so, we develop an extension of our previous analytical model for WiMAX networks, that takes into account a more general scheduling policy than those previously studied (i.e., instantaneous throughput fairness, slot sharing fairness and opportunistic scheduling). We show that this general policy covers the two extreme cases, namely the instantaneous throughput fairness policy and the opportunistic policy, and offers intermediate policies that are good candidates for finding the right trade-off. In order to formulate the decision criterion, we introduce a new performance parameter, the mean throughput obtained by a user depending on his efficiency to use the resource. The model has a closed-form solution, and all performance parameters can be obtained at a click speed. This allows us to carry on dimensioning studies that require several thousands of evaluations, which would not be tractable with any simulation tool.**

I. INTRODUCTION

The evolution of last-mile infrastructure for wired broadband networks faces acute implications such as difficult terrain and high cost-to-serve ratio. Latest developments in wireless domain could not only address these issues but could also complement the existing framework. One of such highly anticipated technologies is WiMAX (Worldwide Interoperability for Microwave Access) based on IEEE standard 802.16. The first operative version of IEEE 802.16 is 802.16-2004 (fixed/nomadic WiMAX) [1]. It was followed by a ratification of mobile WiMAX amendment IEEE 802.16e in 2005 [2]. On the other hand, the consortium WiMAX Forum was found to specify profiles (technology options are chosen among those proposed by the IEEE standard), define an end-to-end architecture (IEEE does not go beyond physical and MAC layer), and certify products (through inter-operability tests).

Some WiMAX networks are already deployed but most operators are still under trial phases. As deployment is coming, the need arises for manufacturers and operators to have fast and efficient tools for network design and performance evaluation. Moreover, no specific scheduling scheme has been recommended in the standard. Indeed, how to efficiently share the precious resource among the users while ensuring fairness between them is by itself a complicated task. Add the very high variability of the radio conditions they can experience (due to their possible mobility, the broadband nature of WiMAX networks, etc.) and the problem becomes even more challenging.

Many scheduling algorithms, developed especially for OFDMA (the technology constituting the PHY layer of WiMAX), have been proposed in the literature [3], [8], [11]–[13] to handle best effort data traffic while providing a good trade-off between an efficient use of the resource and fairness among the users.

[13] presents a scheduler bringing fairness in opportunistic OFDMA systems by taking into account each user's bitrate and delay at each frame. [11] considers a packet scheduling scheme providing fairness among users with the use of a time-utility function as a scheduling urgency factor. In [3] is introduced an optimal solution to the problem of scheduling and resource allocation ensuring fairness through a modified proportional fair scheduling scheme. These three solutions present more or less limited tunable fairness levels, but none of them proposes a way to determine which level should be considered. Besides, they all result in very complex schedulers that must keep track of the constantly changing amount of resources received by each mobile.

On the opposite, [8] and [12] focus on more basic schedulers. However, these two solutions tend to either favour efficiency or fairness. Indeed, in [8] the fairness among user is only considered through a simple embedded round-robin scheme between data connection, while in [12], the efficient use of the resource only intervenes after guaranteeing every users a minimal throughput, whatever their radio conditions. Moreover, these two algorithms stands as unique solutions and, as such, are not adjustable to different radio channel characterizations.

Finally, let us emphasize that all these propositions are based on packet-level simulations that precisely implement system details and scheduling schemes. As a consequence, they correspond to very specific system assumptions.

To overcome these drawbacks, we propose to tackle this scheduling policy design problem from the new angle of analytical modeling. In [4], we developed a novel and generic analytical model able to take into account frame structure, precise slot sharing-based scheduling and channel quality variation of WiMAX systems. Unlike existing models [5], [6], [10], our model is adapted to WiMAX systems' assumptions and is generic enough to integrate any appropriate scheduling policy.

In our previous study [4], we focused on three typical

scheduling policies: the slot sharing fairness, the instantaneous throughput fairness and the opportunistic scheduling. While the instantaneous throughput fairness totally favours fairness between users over an efficient use of the ressource, the opportunistic scheduling does the exact opposite. The slot sharing can be seen as a particular trade-off between these two opposite strategies.

In this paper, we present an extension of our model allowing to easily consider any kind of intermediate scheduling policy, i.e., any kind of trade-off between focusing on efficiency or fairness. Those policies are memoryless (i.e., the amount of resource allocated to a user at a given frame does not depend on the the amount it previously obtained) and, as such, correspond to very simple schedulers, generic enough and easy to implement.

We then propose methods to determine which policy should be used based on radio conditions of users. Two different cases are explored corresponding to two opposite channel assumptions. In the first one, users are assumed to experience very fast changes of their radio conditions (on a frame by frame basis). As a result, all users in active transfer experience similar (good and bad) radio conditions in the same proportions. It will be shown in this case that the opportunistic scheduling is the best policy. In the second case, radio conditions of users in active transfer are assumed to change slowly with respect to the transfer duration. As a result, a user keeps the same radio conditions during its whole transfer. In this case, a compromise must be found to respect a given fairness between users with good and bad conditions, while maintaining an acceptable usage of the resource.

The rest of the paper is organized as follows. Section II presents the analytical model and its extensions. Validation and robustness of the model are discussed in Section III. Section IV finally gives examples of scheduling policy designs.

II. WiMAX Analytical Model

A. Modeling Assumptions

The development of our analytical model is based on several assumptions related to the system, the channel, the traffic and the scheduling algorithm. The assumptions concerning the system, the channel and the traffic, are the same as those already presented, discussed and validated in [4]. Here we recall these assumptions and mostly concentrate on assumptions related to scheduling. Wherever required, related details of WiMAX system are specified. Various notations are also introduced in this section.

A WiMAX time division duplex (TDD) frame comprises of slots that are the smallest unit of resource and which occupies space both in time and frequency domain. A part of the frame is used for overhead (e.g., DL_MAP and UL_MAP) and the rest for user data. The duration T_F of this TDD frame is equal to 5 ms [2].

System assumptions: We consider a single WiMAX cell and focus on the downlink part which is a critical portion of asymmetric data traffic.

1) Overhead in the TDD frame is assumed to be constant and independent of the number of concurrent active mobile station (MS). As a consequence, the total number of slots available for data transmission in the downlink part is constant and will be denoted by N_S.
2) We assume that the number of MS that can simultaneously be in active transfer is not limited. As a consequence, any connection demand will be accepted and no blocking can occur.

One of the important features of IEEE 802.16e is link adaptation: different modulation and coding schemes (MCS) allows a dynamic adaptation of the transmission to the radio conditions. As the number of data subcarriers per slot is the same for all permutation schemes, the number of bits carried by a slot for a given MCS is constant. The selection of appropriate MCS is carried out according to the value of signal to interference plus noise ratio (SINR). In case of outage, i.e., if the SINR is too low, no data can be transmitted without error. We denote the radio channel states as: MCS_k, $1 \leq k \leq K$, where K is the number of MCS. By extension, MCS_0 represents the outage state. The number of bits transmitted per slot by a MS using MCS_k is denoted by m_k. For the particular case of outage, $m_0 = 0$.

Channel assumption: The MCS used by a given MS can change very often because of the high variability of the radio link quality in WiMAX networks. The radio channel may be highly variable (i.e., change from one frame to another) or may vary with some memory (i.e., be maintained during a mean number of frames).

3) We assume that each MS sends a feedback channel estimation on a frame by frame basis, and thus, the base station (BS) can change its MCS every frame. Since we do not make any distinction between users and consider all MS as statistically identical, we associate a probability p_k with each coding scheme MCS_k, and assume that, at each time-step T_F, any MS has a probability p_k to use MCS_k.

As a result, our analytical model only depends upon stationary probabilities of using the different MCS whatever be the radio channel dynamics. We show in Section III through simulation, the robustness of our model against this memoryless channel assumption.

Traffic assumptions: The traffic model is based on the following assumptions.

4) All users have the same traffic characteristics. In addition, we don't consider any QoS differentiation here.
5) We assume that there is a fixed number N of MS that are sharing the available bandwidth of the cell.
6) Each of the N MS is assumed to generate an infinite length ON/OFF elastic traffic. An ON period corresponds to the download of an element (e.g., a web page including all embedded objects). The downloading duration depends on the system load and the radio link quality, so ON periods must be characterized by their size. An OFF period corresponds to the reading time of the last downloaded element, and is independent of the system load. As opposed to ON, OFF periods must then be characterized by their duration.
7) We assume that both ON sizes and OFF durations are

exponentially distributed. We denote by $\bar{x}_{on}$ the average size of ON data volumes (in bits) and by $\bar{t}_{off}$ the average duration of OFF periods (in seconds).

Scheduling assumption: The scheduling algorithm is responsible for allocating radio resources to users. In wireless networks, scheduling may take into account their radio link quality. In previous works [4], we focused on three traditional schemes:

- The slot sharing fairness scheduling equally divides all slots of each frame between all active users that are not in outage.
- The instantaneous throughput fairness scheduling shares the resource in order to provide the same instantaneous throughput to all active users not in outage.
- The opportunistic scheduling gives all the resources to active users having the highest transmission bit rate, i.e., the better MCS.

The two last schemes correspond to two extreme cases. Indeed, while the instantaneous throughput fairness scheduling totally favours fairness between active mobiles over an efficient use of the ressource, the opportunistic scheduling does the exact opposite. Finally, the slot sharing scheduling can be seen as a particular trade-off between these two policies.

In this paper, we provide an extension of our previous analytical model [4] that integrates a more general scheduling policy. Through a parameter γ, this general policy that includes as particular cases the three pre-cited policies, provides a more flexible way to find the right compromise between efficiency and fairness. We assume however that any particular policy issued from this general scheme (and corresponding to a given value of γ), satisfies the following assumption:

8) If there is only one active user (not in outage), the scheduler allocates all the available slots for its transfer.

This assumptions is justified by the fact that, in contrast to some cellular networks (e.g., (E)GPRS), in WiMAX networks, MS do not have limited transmission capabilities (e.g., resulting from hardware considerations). However, one of the QoS parameters considered by the WiMAX standard for traffic classes is the maximum sustained traffic rate (MSTR), which is an upper bound for user throughput. Taking into account MSTR implies the implementation of new throttling scheduling policies that lies outside the range of policies that are considered in this paper. Note that such a throttling policy has been studied in [7].

B. Markovian model

A first attempt for modeling this system would be to develop a multi-dimensional Continuous Time Markov Chain (CTMC). A state $(n_0, ..., n_K)$ of this chain would be a precise description of the current number n_k of MS using coding scheme MCS_k, $0 \leq k \leq K$ (including outage). The derivation of the transitions of such a model is an easy task. However the complexity of the resolution of this model makes it intractable for any realistic value of K. In order to work around the complexity problem, we aggregate the state description of the system into a single dimension n, representing the total number of concurrent active MS, regardless of the MCS they use. The resulting CTMC is thus made of $N + 1$ states as shown in Fig 1 [4].

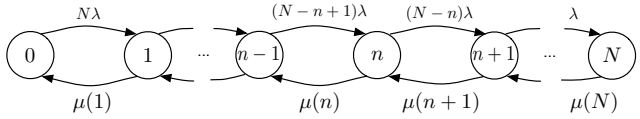


Fig. 1. General CTMC with state-dependent departure rates.

- A transition out of a generic state n to a state $n+1$ occurs when a MS in OFF period starts its transfer. This "arrival" transition corresponds to one MS among the $(N - n)$ in OFF period, ending its reading, and is performed with a rate $(N - n)\lambda$, where λ is defined as the inverse of the average reading time: $\lambda = \frac{1}{\bar{t}_{off}}$.
- A transition out of a generic state n to a state $n - 1$ occurs when a MS in ON period completes its transfer. This "departure" transition is performed with a generic rate $\mu(n)$ corresponding to the total departure rate of the frame when n MS are active.

Obviously, the main difficulty of the model resides in estimating the aggregate departure rates $\mu(n)$. In order to do so, we first express $\mu(n)$ as follows:

$$\mu(n) = \frac{\bar{m}(n)\, N_S}{\bar{x}_{on}\, T_F}, \tag{1}$$

where $\bar{m}(n)$ is the average number of bits transmitted per slot when there are n concurrent active transfers. Obviously, $\bar{m}(n)$ depends on K, the number of MCS, and p_k, $0 \leq k \leq K$, the MCS vector probability. It also strongly depends on n, because the average number of bits per slot must be estimated by considering all possible distributions of the n MS between the $K + 1$ possible MCS (including outage). Finally, it is worthwhile noting that the parameters $\bar{m}(n)$ depend on the scheduling policy, as it defines, at each time-step, the quantity of slots given to each of the n MS with respect to the MCS they use.

In order to provide a generic expression of $\bar{m}(n)$, we first define $x_k(n_0, ..., n_K)$ the proportion of the resource (i.e., of the N_S slots) that is associated to a MS using MCS_k, when the current distribution of the n MS among the $K+1$ coding schemes is $(n_0, ..., n_K)$. The average number of bits per slot, $\bar{m}(n)$, when there are n active users, can then be expressed as follows:

$$\bar{m}(n) = \sum_{\substack{(n_0,...,n_K) = (0,...,0)| \\ n_0 + ... + n_K = n}}^{(n,...,n)} \left(\sum_{k=1}^{K} m_k n_k x_k(n_0, ..., n_K) \right) \cdot P(n_0, ..., n_K), \tag{2}$$

where $P(n_0, ..., n_K)$ is the probability that the current distribution of the n MS among the $K + 1$ coding schemes is $(n_0, ..., n_K)$:

$$P(n_0, ..., n_K) = \binom{n}{n_0, ..., n_K} \prod_{k=0}^{K} p_k^{n_k}. \tag{3}$$

In this relation, $\prod_{k=0}^{K} p_k^{n_k}$ represents the probability of any distribution of the MS such that the number of MS using MSC_k is n_k, and $\binom{n}{n_0,...,n_K}$ is the multinomial coefficient that takes into account all such possibles distributions.

C. Scheduling policy

At this step, all that is left is to derive an expression of the $x_k(n_0, ..., n_K)$ parameters general enough to account for any intermediate policy between the instantaneous throughput fairness and the opportunistic policies. We thus introduce a general parameter γ and express the proportions $x_k(n_0, ..., n_K)$ as follows:

$$x_k(n_0, ..., n_K) = \begin{cases} \dfrac{m_k^\gamma}{\sum_{i=1}^{K} m_i^\gamma \, n_i} & \text{if } k \neq 0 \text{ and } n_k \neq 0 \\ 0 & \text{else} \end{cases} \tag{4}$$

γ represents the relation between the importance of the proportion of the resource granted to a MS and how efficiently it will use it (i.e., the number of bits per slot it can transmit with its current MCS). γ is a real number that can be positive or negative. To a given value of γ corresponds a specific scheduling policy:

- When $\gamma = 0$, each active MS not in outage receives the same portion of the resource whatever its MCS: $x_k(n_0, ..., n_K) = \frac{1}{n-n_0}$ (for any $k \neq 0$). As a result, the corresponding policy corresponds to the slot sharing fairness policy as defined in [4].
- $\gamma > 0$ corresponds to scheduling policies increasing the number of slots allocated to a MS with the capacity of its MCS. These policies clearly favour an efficient use of the ressource over fairness between users to the point that when $\gamma \to +\infty$, only the active users with the very best MCS get to use the resource. This case corresponds to the opportunistic scheduling policy [4]. Note that when γ grows, the corresponding policy rapidly tends towards the opportunistic scheduling. As a result, in Sections 3 and 4, we take $\gamma = 20$ to represent the opportunistic policy.
- $\gamma < 0$, on the contrary, leads to opposite policies, i.e., policies giving more slots to MS with worse MCS, outage excluded, to compensate their lower transmiting capacities. Among those is $\gamma = -1$ that is such that $m_k \, x_k(n_0, ..., n_K) = \frac{1}{\sum_{i=1}^{K} m_i^{-1} n_i} = C$ for all k. Any active MS, whatever the coding scheme it uses, thus receives the same instantaneous throughput (corresponding to $m_k \, x_k(n_0, ..., n_K) \, N_S$ bits per frame). As a result the corresponding policy is the instantaneous throughput fairness policy as defined in [4]. Note that in the remainder of this paper, we don't consider any $\gamma < -1$ as giving better throughputs to MS with lower MCS would lead to nonsensical policies.

Lastly, let us emphasize that, for any of these scheduling policies (i.e., any given γ), the average numbers of bits per slot, $\bar{m}(n)$, rapidly tend to an asymptote as n increases. Indeed, when $n \to +\infty$, the proportion of mobiles using MCS_k tends to p_k, so $n_k \sim p_k \, n$, and

$$\bar{m}(\infty) = \lim_{n \to +\infty} \bar{m}(n) = \frac{\sum_{k=1}^{K} m_k^{\gamma+1} \, p_k}{\sum_{k=1}^{K} m_k^\gamma \, p_k}. \tag{5}$$

Thanks to this asymptotical behavior, we can avoid the costly computation of the $\bar{m}(n)$ for large values of n (e.g., by replacing, the exact values by the asymptote as soon as they become close enough).

D. Performance parameters

The steady-state probabilities $\pi(n)$ can easily be derived from the birth-and-death structure of the Markov chain (depicted in Fig. 1):

$$\pi(n) = \frac{N!}{(N-n)!} \frac{T_F^n \, \rho^n}{N_S^n \prod_{i=1}^{n} \bar{m}(i)} \, \pi(0), \tag{6}$$

where ρ is given by relation (7) and plays a role equivalent to the "traffic intensity" of Erlang laws [9], and $\pi(0)$ is obtained by normalization.

$$\rho = \frac{\bar{t}_{on}}{\bar{t}_{off}} \tag{7}$$

The performance parameters of this system can be derived from the steady-state probabilities as follows.

As a consequence of assumption 8, the average utilization $\bar{U}$ of the TDD frame is one as long as there is at least one active mobile that is not in outage:

$$\bar{U} = \sum_{n=1}^{N} (1 - p_0^n) \pi(n). \tag{8}$$

The average number of active users $\bar{Q}$ is expressed as:

$$\bar{Q} = \sum_{n=1}^{N} n \, \pi(n). \tag{9}$$

The mean number of departures $\bar{D}$ (MS completing their transfer) by unit of time, is given by:

$$\bar{D} = \sum_{n=1}^{N} \pi(n) \, \mu(n). \tag{10}$$

From Little's law, we can derive the average duration $\bar{t}_{on}$ of an ON period (duration of an active transfer):

$$\bar{t}_{on} = \frac{\bar{Q}}{\bar{D}}, \tag{11}$$

and deduce the average throughput $\bar{X}$ obtained by each MS in active transfer:

$$\bar{X} = \frac{\bar{x}_{on}}{\bar{t}_{on}}. \tag{12}$$

Finally, we express the average throughput $\bar{X}_k$ obtained by each MS in active transfer while using MCS_k by:

$$\bar{X}_k = \frac{\bar{B}_k}{\bar{Q}_k \, T_F}, \tag{13}$$

where $\bar{B}_k$ is the mean number of bits per frame transmitted by all MS using MCS_k and $\bar{Q}_k$ is the average number of active

MS using MCS_k. It is easy to demonstrate that $\bar{Q}_k$ is nothing but a proportion p_k of $\bar{Q}$:

$$\bar{Q}_k = \sum_{n=1}^{N} \pi(n) \sum_{\substack{(n_0,...,n_K)=(0,...,0)| \\ n_0+...+n_K=n}}^{(n,...,n)} n_k \, P(n_0,...,n_K)$$

$$= \sum_{n=1}^{N} \pi(n) \, n \, p_k = p_k \, \bar{Q}, \quad (14)$$

and $\bar{B}_k$ has the following expression:

$$\bar{B}_k = \sum_{n=1}^{N} m_k \, \bar{x}_k(n) \, N_S \, \pi(n), \quad (15)$$

where $\bar{x}_k(n)$ is the average portion of the frame granted to MS using MCS_k when there are n active MS:

$$\bar{x}_k(n) = \sum_{\substack{(n_0,...,n_K)=(0,...,0)| \\ n_0+...+n_K=n}}^{(n,...,n)} n_k \, x_k(n_0,...,n_K) \cdot P(n_0,...,n_K). \quad (16)$$

Observe that just like for the computation of the $\bar{m}(n)$, we can easily avoid the calculation of the $\bar{x}_k(n)$ for large values of n by considering the following asymptote:

$$\bar{x}_k(\infty) = \lim_{n \to +\infty} \bar{x}_k(n) = \frac{m_k^\gamma \, p_k}{\sum_{i=1}^{K} m_i^\gamma \, p_i}. \quad (17)$$

Lastly, as demonstrated in Appendix I, the global throughput $\bar{X}$ can be related to the specific throughputs $\bar{X}_k$ by the following straightforward relation:

$$\bar{X} = \sum_{k=1}^{K} p_k \, \bar{X}_k, \quad (18)$$

since a MS has a probability p_k to use MCS_k and, as a consequence, to achieve an average throughput $\bar{X}_k$.

III. Validation

In this section we discuss the validation and robustness of our analytical model through extensive simulations. Validation refers to comparing analytical and simulation results under the same traffic and channel assumptions, while robustness means considering simulations with more realistic assumptions that are not explicitly taken into account in the model. The validation study mainly focuses on the extensions of the model that have not been previously validated in [4], i.e., the accuracy of the model for any intermediate scheduling policy and for the new performance parameters $\bar{X}_k$. The robustness study concentrates on showing the robustness of our analytical model with respect to the channel assumptions, as it becomes a key point of our study when exploring the trade-off between fairness and efficiency under low variability of radio conditions. In the following, we first present the details of the discrete-event simulator we have developed to validate our model. Then we provide the validation study and the robustness study.

A. Simulation Models

Here we present our discrete-event simulator and the assumptions it stands on.

System Parameters: The number of slots in downlink, N_S, depends on the system bandwidth, frame duration, downlink/uplink ratio, subcarrier permutation (PUSC, FUSC or AMC) and the protocol overhead (preamble, FCH, maps). So, by assuming a system bandwidth of 10 MHz, a TDD frame duration of $T_F = 5$ ms, a constant downlink/uplink ratio of 2/3, a PUSC subcarrier permutation and, for the sake of simplicity, a protocol overhead of fixed length (2 symbols), we obtain $N_S = 450$ slots.

Channel Models: In simulation, the wireless channel between the BS and a MS can feature a memory. This means that the state of a channel at a given frame, which determines the coding rate MCS to use for the MS, can influence the state of this channel for the next coming frame. This memory is simply taken into account through a real-valued parameter a between 0 and 1. a defines the probability that an active MS maintains the same MCS for the next trame. Thus, an active MS keeps its MCS for a certain duration, referred to as the time of coherence, whose mean is $\bar{t}_{coh} = 1/(1-a)$. With $a = 0$, the channel is memoryless, i.e. MCS are independently drawn from frame to frame for each user. On the other hand, with $a = 1$, a MS will never change its MCS. For the simulation results below, we have considered values of a equal to 0, 0.9 and 0.99, which correspond to average times of coherence of respectively 1, 10 and 100 frames. Note that $\bar{t}_{coh} = 100$ trames means that a MS only changes its MCS, on average, every 500 ms.

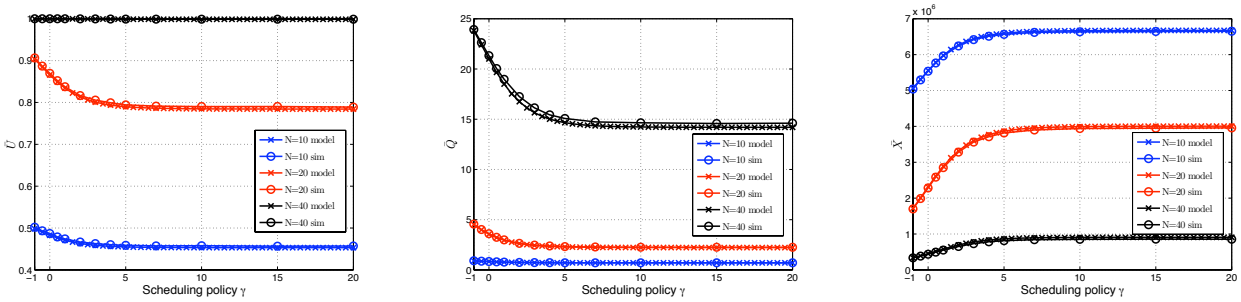
Let us remind that the analytical model only takes into account the channel variability through the average probabilities p_k of using MCS_k, and makes no difference between memory and memoryless channels. It is thus of primary importance to validate the robustness of the model with respect to this channel assumption.

Traffic Model: The traffic model used in the simulator is the same as the one used in the analytical model and presented in Section II-A, i.e., an infinite-length elastic ON/OFF traffic, where ON size and OFF duration are assumed to be exponentially distributed with respective means $\bar{x}_{on}$ an $\bar{t}_{off}$. Note that previous works [4] have shown the robustness of the analytical model to other traffic characterizations (e.g. Pareto distribution for ON size).

Scheduling Model: While our analytical model only takes into account bits per slot averages, $\bar{m}(n)$, in the computation of the departure rates $\mu(n)$, the simulator implements a complete centralized scheduler that allocates slots to active users on a frame by frame basis. To do so, at each frame, the simulator determines the number of slots allocated to each MS based on its current MCS, following relation (4) with γ set to the considered scheduling policy.

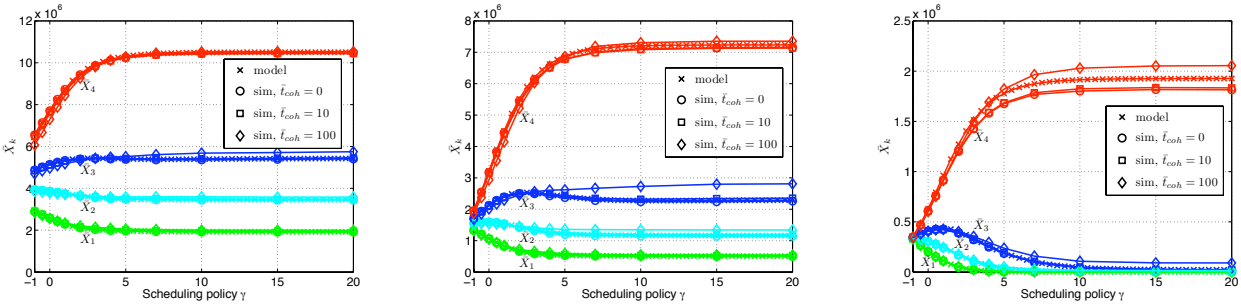
B. Simulation Results

We now compare the results obtained through our analytical model with simulations. Table I summarizes the system and traffic parameters and Table II shows the wireless channel



(a) Average utilization, $\bar{U}$ (b) Average number of active users, $\bar{Q}$ (c) Average throughput per user, $\bar{X}$

Fig. 2. Customary performance parameters with different scheduling policies, increasing levels of traffic load and a memoryless wireless channel.



(a) Low traffic load: $N = 10$ (b) Moderate traffic load: $N = 20$ (c) Heavy traffic load: $N = 40$

Fig. 3. Average throughput of a user, $\bar{X}_k$, provided he uses MCS_k with different scheduling policies, levels of traffic load, and $\bar{t}_{coh}$.

parameters (i.e., the MCS, outage included, and their associated stationary probabilities) considered in the following simulations. All the results are presented for a wide range of scheduling policies, from $\gamma = -1$ (instantaneous throughput fairness) to $\gamma = 20$ (opportunistic scheduling), and three levels of traffic load: low load ($N = 10$ MS), moderate load ($N = 20$ MS) and high load ($N = 40$ MS).

TABLE I
SYSTEM AND TRAFFIC PARAMETERS.

Parameter	Value
Number of slots per trame, N_S	450 slots
Duration of a trame, T_F	5 ms
Scheduling policy, γ	-1 to 20
Number of MS in the cell, N	10, 20 or 40 MS
Average size of ON data volumes, $\bar{x}_{on}$	3 Mbits
Average duration of OFF periods, $\bar{t}_{off}$	6 s

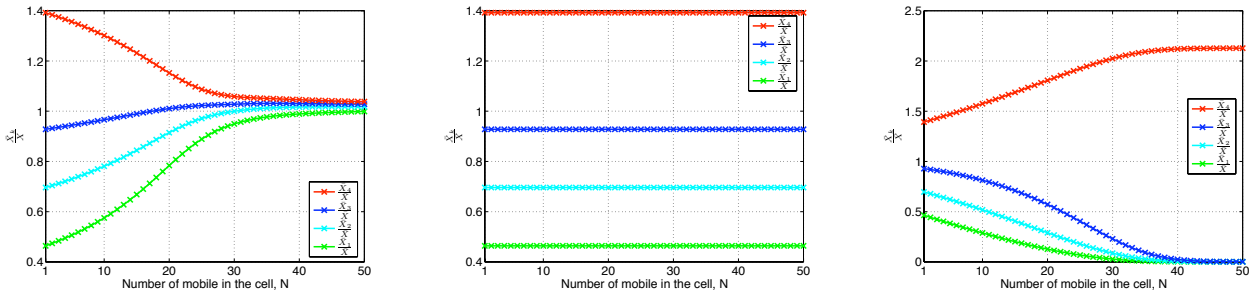
TABLE II
CHANNEL PARAMETERS.

Channel state $\{0, \ldots, K\}$	MCS and outage	Bits per slot m_k	Probability of use p_k
0	Outage	$m_0 = 0$	0.02
1	QPSK-1/2	$m_1 = 48$	0.12
2	QPSK-3/4	$m_2 = 72$	0.31
3	16QAM-1/2	$m_3 = 96$	0.08
4	16QAM-3/4	$m_4 = 144$	0.47

Validation: We first compare the results obtained by our model and those delivered by simulations with a memoryless channel ($\bar{t}_{coh} = 0$ frame). Fig. 2 illustrates the good accuracy of our analytical model for evaluating customary performance parameters, $\bar{U}, \bar{Q}, \bar{X}$ (obtained from relations 8 to 12), for any level of traffic loads and any intermediate scheduling policy. In the same conditions, analytical and simulation results are very close for $\bar{X}_k$ as well, as shown by Fig. 3. This enables us to conclude on the high accuracy of our analytical computation of $\bar{X}_k$ (relation 13).

Robustness: We then confront the analytical model to simulations considering radio channels with memory. Fig. 3 also presents the simulation results obtained for two coherence times, $\bar{t}_{coh} = 10$ and $\bar{t}_{coh} = 100$ frames, corresponding to moderate and high channel memory. When $\bar{t}_{coh} = 10$ frames, these results match the analytical model very closely. And, when $\bar{t}_{coh} = 100$ frames, the differences we can observe remain small, generally less than 8%. In both cases, these results establish the high accuracy of the analytical model even though our model only takes into account the stationary probabilities of the MCS. This tends to show that the channel information is almost completely included in the stationary probabilities of the MCS and, as such, that only considering these probabilities is sufficient to accurately model any channel with memory.

Throughout this section, the extensions of our analytical model have been validated thanks to the simulator. First, we showed that our analytical model is consistent for any intermediate policy. Second, we saw that the analytical computation of the $\bar{X}_k$ is fair for any level of traffic load. We then established that it still holds when considering a wireless channel integrating memory. Indeed, for coherence times $\bar{t}_{coh}$, ranging from 0 to 100 frames, the results indicate that the



(a) Instantaneous throughput fairness (b) Slot sharing fairness (c) Opportunist

Fig. 4. Evolution of the $\frac{\bar{X}_k}{\bar{X}}$ ratios when the traffic load increases of three typical scheduling policies.

differences between the analytical model and simulations are maintained at very low levels. Note finally that simulation results have required several hours of computing whereas analytical results were obtained instantaneously.

IV. Dimensioning

In this section we try to answer the fundamental question of this paper: how to find the right scheduling policy that achieves the best compromise between fairness and efficiency? We investigate this problem in two extreme cases corresponding to high and low variability of radio conditions of users. In both cases we however assume that all users experience statistically the same radio conditions, and, as a result, have the same average probabilities p_k of using coding schemes MCS_k (including outage) during their transfer.

A. Highly variable radio conditions

In this subsection, users are assumed to experience very fast changes of their radio conditions. At the limit, the current MCS of a MS do not affect his forthcoming MCS. As a result, any user has the same probability of getting good or bad conditions at any frame, regardless of its previous radio conditions. This corresponds to the memoryless channel assumption (as defined in Section III). In this highly variable radio conditions scenario, there is no need to impose fairness between users. Indeed, because of the high variability of their radio conditions, all users will change MCS very frequently, and, during their transfer, will thus use the different MCS_k in the same proportions.

Fig. 2 shows the attained performance parameters in such radio conditions (system and traffic parameters are given in Tables I and II). $\bar{U}$, $\bar{Q}$ and $\bar{X}$ are evaluated for several scheduling policies, ranging from the instantaneous fairness (i.e., $\gamma = -1$) to the opportunistic scheduling (i.e., $\gamma = 20$). As can be seen in Fig. 2(b), the mean number of users in active transfer, $\bar{Q}$, decreases significantly as γ increases. As an example, in the case where there are $N = 40$ MS in the cell, $\bar{Q}$ varies from about 24 for $\gamma = -1$ to only 15 for $\gamma = 20$. As the average number of MS that are sharing the radio resources decreases, the average instantaneous throughput obtained by each MS in active transfer, $\bar{X}$, increases with γ (as can be seen on Fig. 2(c)).

Overall, these results clearly show that the opportunistic policy guarantees the most efficient use of the channel resource. Thus, this scheduling policy represents the best choice to implement in the scheduler, as soon as there is no need to ensure fairness among users, e.g., when they experience the same highly variable radio conditions.

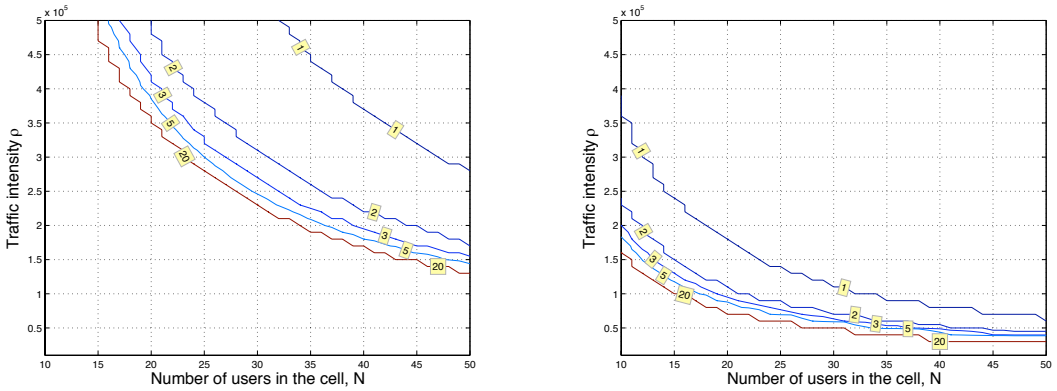
B. Lowly variable radio conditions

We now concentrate on the opposite scenario and assume that radio conditions of users in active transfer change slowly with respect to the transfer duration. We can thus consider that a MS keeps the same radio conditions during the whole duration of its transfer (i.e., uses the same MCS during an ON period). Note that this corresponds to a memory channel with a long enough coherence time (as defined in Section III). In this lowly variable radio conditions scenario, it becomes crucial to guarantee a certain degree of fairness between users by the way of the scheduling policy. Indeed, a mobile can now be stuck with the same bad MCS for a long time and completely deprived of the resource if the policy only serves the MS with better MCS.

To understand how to evaluate fairness between users, we first propose to look into the $\frac{\bar{X}_k}{\bar{X}}$ ratios ($k \neq 0$). Fig. 4 shows the evolution of these ratios when the number of users in the cell, N, goes from 1 to 50 (i.e., when the traffic load increases), for the three usual scheduling policies: instantaneous throughput fairness, slot sharing fairness and opportunistic. (System and traffic parameters corresponding to this scenario are still the ones given in Tables I and II.)

Fig. 4(a) corresponds to the instantaneous throughput fairness policy. We can see that all ratios tend to 1 when the traffic load increases, i.e., that the average throughputs $\bar{X}_k$ obtained by an active MS using MCS_k tend to be the same for all MCS (excluding outage). Note that this is exactly the aim of the instantaneous throughput fairness policy, i.e., to offer the same instantaneous throughput to all MS not in outage.

The $\frac{\bar{X}_k}{\bar{X}}$ ratios when considering the slot sharing fairness policy are illustrated by Fig. 4(b). The gaps between ratios remain constant as each active mobile not in outage receives the same number of slots and uses them with a constant efficiency determined by its MCS. Moreover, we can see that these ratios do not depend on the traffic load. This result is formally demonstrated in Appendix II.



(a) $\alpha = 0.2$ (b) $\alpha = 0.4$

Fig. 5. Dimensioning γ to obtain the most efficient policy guaranteeing $\frac{\bar{X}_k}{\bar{X}} \geq \alpha$ for all $k \neq 0$.

Lastly, Fig. 4(c) depicts the $\frac{\bar{X}_k}{\bar{X}}$ ratios with an opportunistic scheduling. As the traffic load increases, so does the probability that there is at least one active user in the cell with the best MCS (i.e., MCS_K). Since the opportunistic policy only allocates slots to active users having the best current MCS, MS using MCS_K fully use the resource, whereas MS using other MCS see their throughput converging to 0, as shown by the figure.

To sum up, if the policy only focuses on guaranteeing fairness between users (instantaneous throughput fairness), all the $\frac{\bar{X}_k}{\bar{X}}$ ratios tend to the same value when the traffic load increases. On the contrary, if the focus is on the efficient use of the resource (opportunistic), all the ratios, except $\frac{\bar{X}_K}{\bar{X}}$, tend to 0 as only users of MCS_K (the best MCS) get to use the resource.

In order to guarantee a degree α of fairness between users, we propose to only consider the set of scheduling policies leading to $\frac{\bar{X}_k}{\bar{X}} \geq \alpha$, for all $k \neq 0$, where α is an input parameter that has to be chosen between 0 (no fairness) and 1 (maximum fairness). As an example, for $\alpha = 0.2$, the set will consist of all policies that insure an instantaneous throughput for MS using MCS_k that at least equals 20% of the global average throughput, for any k ($k \neq 0$). Then, from this set of policies, we select the one associated with the greatest value of γ, since it corresponds to the most efficient of the set, and as such leads to the best performances.

Following these rules, Figure 5 allows to dimension the scheduling policy for any given traffic load defined by the couple (N, ρ) (where N is the number of MS in the cell and ρ is the traffic intensity given by relation 7), and for two different levels of fairness: $\alpha = 0.2$ and $\alpha = 0.4$. To obtain the graphs constituting this Figure, we first drew 3-dimensional surfaces where the recommanded policy γ is function of the traffic load (N, ρ). Then, we cut out the resulting surfaces into arbitrarily chosen level lines (here corresponding to values of γ of 1, 2, 3, 5 and 20) and projected these lines in a 2-dimensional plane. To use these abacus, we select the point (N, ρ) corresponding to the target traffic load on the graph associated with the level of fairness α we want to garanty. Finally, we just have to choose the value of γ corresponding to the closest line passing above the given point. The corresponding policy will be the one that satisfies the chosen level of fairness while making best use of the resource, i.e., providing the best performance.

As an example, let us consider a cell with $N = 20$ users, each generating a traffic charaterized by $\bar{x}_{on} = 1$ Mbits and $\bar{t}_{off} = 10$ s, i.e, a traffic of intensity $\rho = \frac{10^6}{10} = 10^5$. If we only want to guarantee a low level of fairness $\alpha = 0.2$ then Fig. 5(a) indicates that any value of γ is suitable. The opportunistic policy ($\gamma = 20$) should then be chosen. However, if we decide on a more restrictive level of fairness $\alpha = 0.4$, Fig. 5(b) states that the value of γ shall not exceed 2. As a consequence, in this specific case, the policy associated with $\gamma = 2$ is the best choice.

Also, note that in Fig. 5, if there is no line above the considered point (top right corner of the graphs), then the recommended policy is $\gamma = 0$. Indeed, given the channel parameters, the $\frac{\bar{X}_k}{\bar{X}}$ ratios for all $k \neq 0$, are always greater than $\alpha = 0.2$ and $\alpha = 0.4$ when $\gamma = 0$ whatever the traffic load, as shown in Fig. 4(b).

Finally, let us emphasize that these graphs were obtained through several thousands resolutions of our model almost instantaneously, whereas they would have required several months (years?) of computing by the means of simulations.

V. Conclusion

While WiMAX is increasingly being deployed, no specific scheduling scheme has been yet recommended in the standard. In this paper, we tackle the problem of deciding on the best policy to implement by considering a relatively simple analytical model that integrates, by means of a real-valued parameter γ, any memoryless scheduling policy, ranging from instantaneous throughput fairness ($\gamma = -1$) to opportunistic scheduling ($\gamma = \infty$). All the performance parameters are derived from closed-form expressions at a click speed. Among them, a new parameter has been defined: $\bar{X}_k$, the instantaneous throughput obtained by users conditioned by their efficiency to use the radio resource (i.e., conditioned by the fact that they currently use a given coding scheme MCS_k). We validated

our model thanks to a home-made discrete-event simulator that implements the parametric policies on a frame by frame basis and integrates a more realistic channel model, and as such also validated the robustness of our model with respect to the channel assumptions. Then, we rely on the analytical model to investigate how to choose the "right" policy in two extreme cases. If users experience very fast changes of their radio conditions, it is shown that the opportunistic policy represents the best choice. On the other hand, when the channel variability is low, we propose an efficient way to decide of the intermediate scheduling policy that yields to the right balance between fairness and efficiency. This last point involves the definition of a criterion based on $\bar{X}_k$, and is performed thanks to dimensioning abacus that have been obtained almost instantaneously by invoking thousands resolutions of our analytical model.

APPENDIX I
DEMONSTRATION OF RELATION 18

Starting from the expression of $\bar{X}$ (relation 12), $\bar{X} = \frac{\bar{x}_{on}}{\bar{t}_{on}}$, we replace $\bar{t}_{on}$ then $\bar{D}$ by their expressions (relations 11) and 10):

$$\bar{X} = \frac{\bar{x}_{on} \sum_{n=1}^{N} \mu(n)\,\pi(n)}{\bar{Q}}$$

Then we use the expression (1) of $\mu(n)$:

$$\begin{aligned}
\bar{X} &= \frac{\bar{x}_{on} \sum_{n=1}^{N} \frac{\bar{m}(n)\,N_S}{\bar{x}_{on}\,T_F}\,\pi(n)}{\bar{Q}} \\
&= \frac{\sum_{n=1}^{N}\sum_{k=1}^{K} m_k\,\bar{x}_k(n)\,N_S\,\pi(n)}{\bar{Q}\,T_F} \\
&= \sum_{k=1}^{K} \frac{\sum_{n=1}^{N} m_k\,\bar{x}_k(n)\,N_S\,\pi(n)}{\bar{Q}\,T_F} \\
&= \sum_{k=1}^{K} \frac{p_k \sum_{n=1}^{N} m_k\,\bar{x}_k(n)\,N_S\,\pi(n)}{p_k\,\bar{Q}\,T_F},
\end{aligned}$$

and identify $\bar{B}_k$ and $\bar{Q}_k$ (relations 15 and 14) in this expression:

$$\bar{X} = \sum_{k=1}^{K} \frac{p_k\,\bar{B}_k}{\bar{Q}_k\,T_F}.$$

Finally, we identify the expression of $\bar{X}_k$ (relation 13) and demonstrate the result:

$$\bar{X} = \sum_{k=1}^{K} p_k\,\bar{X}_k.$$

APPENDIX II
EXPRESSION OF THE $\frac{\bar{X}_k}{\bar{X}}$ RATIOS WHEN $\gamma = 0$

From relations 18, 13 and 14, we obtain the $\frac{\bar{X}_k}{\bar{X}}$ ratios:

$$\begin{aligned}
\frac{\bar{X}_k}{\bar{X}} &= \frac{\bar{X}_k}{\sum_{i=1}^{K} p_i\,\bar{X}_i} = \frac{\frac{\bar{B}_k}{\bar{Q}_k\,T_F}}{\sum_{i=1}^{K} p_i\,\frac{\bar{B}_i}{\bar{Q}_i\,T_F}} = \frac{\frac{\bar{B}_k}{p_k\,\bar{Q}\,T_F}}{\sum_{i=1}^{K} \frac{\bar{B}_i}{\bar{Q}\,T_F}} \\
&= \frac{\bar{B}_k}{p_k \sum_{i=1}^{K} \bar{B}_i}.
\end{aligned} \tag{19}$$

$\bar{B}_k$, the mean number of bits per frame transmitted by all the mobiles using MCS_k (relation 15), can also be expressed as:

$$\bar{B}_k = \bar{Q}_k\,\bar{B}_k^1,$$

where $\bar{B}_k^1$ is the mean number of bits per frame transmitted by only one MS using MCS_k:

$$\bar{B}_k^1 = \sum_{n=1}^{N} m_k\,N_S\,\bar{x}_k^1(n)\,\pi(n).$$

In this expression, $\bar{x}_k^1(n)$ is the average portion of the frame granted to one MS using MCS_k when there are n active MS:

$$\bar{x}_k^1(n) = \sum_{\substack{(n_0,\ldots,n_K)=(0,\ldots,0)| \\ n_0+\ldots+n_K=n}}^{(n,\ldots,n)} x_k(n_0,\ldots,n_K)P(n_0,\ldots,n_K).$$

The slot sharing fairness policy ($\gamma = 0$) gives to all active MS not in outage the same portion of the resource whatever their MCS: $x_k(n_0,\ldots,n_K) = \frac{1}{n-n_0}$ (for any $k \neq 0$).

Thus, when $\gamma = 0$ and $k \neq 0$, the $\bar{x}_k^1(n)$ become independent of k:

$$\bar{x}_k^1(n) = \sum_{\substack{(n_0,\ldots,n_K)=(0,\ldots,0)| \\ n_0+\ldots+n_K=n}}^{(n,\ldots,n)} \frac{P(n_0,\ldots,n_K)}{n-n_0} = \bar{x}^1(n),$$

and the $\bar{B}_k$ can be expressed as:

$$\bar{B}_k = \bar{Q}_k\,m_k\,N_S \sum_{n=1}^{N} \bar{x}^1(n)\,\pi(n).$$

Then, by using this expression in relation 19, we obtain:

$$\begin{aligned}
\frac{\bar{X}_k}{\bar{X}} &= \frac{\bar{Q}_k\,m_k\,N_S(\sum_{n=1}^{N} \bar{x}^1(n)\,\pi(n))}{p_k \sum_{i=1}^{K} \bar{Q}_i\,m_i\,N_S(\sum_{n=1}^{N} \bar{x}^1(n)\,\pi(n))} \\
&= \frac{\bar{Q}_k\,m_k}{p_k \sum_{i=1}^{K} \bar{Q}_i\,m_i} = \frac{p_k\,\bar{Q}\,m_k}{p_k \sum_{i=1}^{K} p_i\,\bar{Q}\,m_i} \\
&= \frac{m_k}{\sum_{i=1}^{K} p_i\,m_i}.
\end{aligned}$$

From this last result, we conclude that when considering the slot sharing fairness policy, the $\frac{\bar{X}_k}{\bar{X}}$ ratios only depend on the MCS_k and the probabilities p_k. As a consequence, in this particular case, they are not affected by the traffic load.

REFERENCES

[1] IEEE Standard for local and metropolitan area networks - Part 16: Air Interface for Fixed Broadband Wireless Access Systems, 2004.

[2] Draft IEEE std 802.16e/D9. IEEE Standard for local and metropolitan area networks - Part 16: Air Interface for Fixed Broadband Wireless Access Systems., 2005.

[3] R. Agrawal, R. Berry, J. Huang, and V. Subramanian. Optimal Scheduling for OFDMA Systems. In *Signals, Systems and Computers, 2006. ACSSC '06. Fortieth Asilomar Conference*, November 2006.

[4] B. Baynat, G. Nogueira, M. Maqbool, and M. Coupechoux. An Efficient Analytical Model for the Dimensioning of WiMAX Networks. In *IFIP Networking 2009*.

[5] T. Bonald and A. Proutiere. Wireless downlink channels: User performance and cell dimensioning. In *ACM Mobicom*, 2003.

[6] S. Borst. User-level performance of channel-aware scheduling algorithms in wireless data networks. In *IEEE Infocom*, 2003.

[7] S. Doirieux, B. Baynat, M. Maqbool, and M. Coupechoux. An Analytical Model for WiMAX Networks with Multiple Traffic Profiles and Throttling Policy. In *Proc. of the 7th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks*.

[8] M. Einhaus and O. Klein. Performance Evaluation of a basic OFDMA Scheduling Algorithm for Packet Data Transmissions. In *Proc. of the 11th IEEE Symposium on Computers and Communications*, June 2006.

[9] T. O. Engset. On the calculation of switches in an automatic telephone system. In *Tore Olaus Engset: The man behind the formula*, 1998.

[10] S. Liu and J. Virtamo. Performance Analysis of Wireless Data Systems with a Finite Population of Mobile Users. In *19th ITC*, 2005.

[11] S. Ryu, B. Ryu, H. Seo, M. Shin, and S. Park. Wireless Packet Scheduling Algorithm for OFDMA System based on Time-utility and Channel State. In *Vehicular Technology Conference (VTC) IEEE 61st*, June 2005.

[12] V. Singh and V. Sharma. Efficient and Fair Scheduling of Uplink and Downlink in IEEE 802.16 OFDMA Networks. In *Wireless Communications and Networking Conference, 2006. WCNC 2006. IEEE*, April 2006.

[13] P. Svedman, S. Wilson, L. Cimini, and B. Ottersten. Opportunistic Beamforming and Scheduling for OFDMA Systems. In *Communications, IEEE Transactions on, Volume 55, Issue 5*, May 2007.