Performance Evaluation and Dimensioning of WiMAX

G. Nogueira, B. Baynat, M. Maqbool, M. Coupechoux

Book Title: XXXXXXXXXXXXXXXXXXXXXXXXXXXX

Editors

July 1, 2008

Contents

Chapter 1

Performance Evaluation and Dimensioning of WiMAX

Georges Nogueira, Bruno Baynat (UPMC - Paris6)
and Masood Maqbool, Marceau Coupechoux (ENST - Paris)

1.1 Abstract

This chapter tackles the challenging task of performance evaluation and dimensioning of WiMAX networks. It provides a simple analytical model which is able to take into account the effects of elastic traffic, radio channel variations and scheduling policy. Compared to packet-level simulation based evaluations, our model instantaneously delivers the dimensioning parameters necessary for the deployment of a WiMAX network. Compared to existing analytical solutions, we derive closed-form expressions for all performance metrics. We compare the results obtained through analytical model with those of simulations. We show that our analytical model is not only accurate but also robust with respect to the modeling assumptions. Finally, the quick results produced through our analytical tool allows to carry out dimensioning analyses that otherwise require several thousands of evaluations, which would not be tractable with any simulation tool.

1.2 Introduction

In recent years, the demand for broadband access has increased substantially. To date, most of the deployed broadband networks are wired ones. The evolution of last-mile infrastructure for wired networks faces acute implications such as difficult terrain and high cost-to-serve ratio. Latest developments in the 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 the standard IEEE 802.16. The first operative version of IEEE 802.16 is 802.16-2004 (fixed/nomadic WiMAX) [2]. It was followed by a ratification of amendment IEEE 802.16e (mobile WiMAX) in 2005 [3]. A new standard, IEEE 802.16m, is currently under definition for providing even higher efficiency. Besides, the consortium WiMAX Forum was founded 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 certificate products (through inter-operability tests). Some WiMAX networks are already deployed but most operators are still under trial phases. As deployment is approaching, the need arises for manufacturers and operators to have fast and efficient tools for network design and performance evaluation. In this chapter, we develop a simple and accurate analytical model that allows to rapidly derive the capacity parameters such as throughput per user, channel utilization or mean number of active users for different scheduling policies.

1.3 WiMAX performance evaluation

Literature on WiMAX performance evaluation is mainly constituted of two sets of papers. One set discusses detailed packet-level simulations that precisely implement system details and scheduling schemes while the other one focuses on analytical models and optimizations.

In the former set, Lee et al. [18] have presented a simulation based performance analysis for three different classes of services proposed in IEEE 802.16e: UGS (Unsolicited Grant Service), rtPS (real time Polling Service) and ertPS (extended real time Polling Service).

The application in context was VoIP (Voice over IP). The authors concluded that ertPS could accommodate more voice calls while satisfying the constraint of minimum packet delay. The performance analysis was focused on MAC layer. Cicconetti et al. [8] analyzed the performance of IEEE 802.16 system in providing multiple services (i.e. web and VoIP). The authors have investigated QoS support mechanisms of the standard in conjunction with classical scheduling schemes like DRR (Deficit Round Robin) or WRR (Weighted Round Robin). See, for example, also [15, 24, 26] on the same subject.

Among the second set of papers, authors of [27] propose an analytical model for studying the random access scheme of IEEE 802.16d. Authors consider a perfect radio channel, that is certainly unrealistic in WiMAX networks. This model finally allows to configure an extension of exponential backoff algorithm applied to IEEE 802.16d. For the same standard, Singh and Sharma [25] consider UGS users and present a scheduling algorithm to minimize the global unsatisfaction of these circuit switched class of users. Linear programming is used to formalize the problem and heuristic algorithms are proposed. Finally, Niyato and Hossain [21] formulate the bandwidth allocation of multiple services with different QoS requirements by using linear programming. They also propose performance analysis, first at connection level, and then, at packet level. In the former case, variations of the radio channel are, however, not taken into account. In the latter case, the computation of performance measures relies on a multidimensional Markovian model that requires numerical resolution.

Not specific to WiMAX systems, generic analytical models for performance evaluation of cellular networks with varying channel conditions have been proposed in [12, 13, 19]. The models presented in these articles are mostly based on multi-class processor-sharing queues with each class corresponding to users having similar radio conditions and subsequently equal data rates. The variability of radio channel conditions at flow level is taken into account by integrating a propagation model, classical mobility models or, in some cases, a spatial distribution of users in a cell. For example, [12] and [13] consider a spatial distribution of users in a cell made of constant capacity rings obtained through a classical Rayleigh fading distribution. In order to use classical PS-queues results, these papers consider implicitly that users can only switch class between two successive data transfers. However, as highlighted in the next section, in WiMAX systems, radio conditions and thus data rates of a particular

user can change frequently during a data transfer. In addition, capacity of a WiMAX cell may vary as a result of varying radio conditions of users. As a consequence, any PS, DPS (discriminatory PS) or even GPS (generalized PS) queue is not appropriate for modeling these channel variations.

In this chapter, we develop a novel and generic analytical model that takes into account frame structure, precise slot sharing-based scheduling and channel quality variation of WiMAX systems. Unlike existing models [12, 13, 19], our model is adapted to WiMAX systems' assumptions and is generic enough to integrate any appropriate scheduling policy. Here, we consider three classical policies: *slot sharing fairness*, *instantaneous throughput fairness*, and *opportunistic*. For each of them, we develop closed-form expressions for all performance metrics. Moreover, our approach makes it possible to take into account the so-called "outage" situation. A given user experiences an outage, if at a given time its radio conditions are so bad that it cannot transfer any data and is thus not scheduled. Once again, classical PS-like queues are not appropriate to model this feature.

1.4 WiMAX System Description

In this section, we briefly present the WiMAX system details needed to understand the proposed analytical model. Although the analysis is also valid for fixed WiMAX, we focus on mobile WiMAX, which is based on standard IEEE 802.16e and SOFDMA (Scalable Orthogonal Frequency Division Multiple Access) physical layer. In particular, the WiMAX frame structure, the notion of radio resource (slot), the access technique, and the different Modulation and Coding Scheme (MCS) are presented.

The PHY layer of WiMAX is based on OFDMA. OFDM splits the available spectrum into a number of parallel orthogonal narrowband subcarriers, grouped into multiple subchannels. Radio resources are thus available in terms of OFDM symbols (time domain) and subchannels (frequency domain) providing a time-frequency multiple access technique [16]. In IEEE 802.16e, possible system bandwidths are 20, 10, 5 and 1.25 MHz with associated FFT (Fast Fourier Transform) sizes of 2048, 1024, 512 and 128 respectively [1]. The total number of

subchannels depends on the subcarrier permutation, i.e., the way subcarriers are grouped together. Two main methods mentioned in [1] are: distributed and adjacent subcarrier permutations. Full usage of subchannels (FUSC) and Partial usage of subchannels (PUSC) are examples of distributed subcarrier permutations; they take advantage of channel diversity among subchannels. Adaptive modulation and coding (AMC) is a type of adjacent subcarrier permutation which allows an opportunistic use of the channel.

IEEE 802.16e has specified time division duplex (TDD) as duplexing technique. The ratio of downlink (DL) to uplink (UL) has been left open in the standard. WiMAX Forum has specified a duration of TDD frame of 5 ms. An example of a WiMAX TDD frame is shown in Fig. 1.1. It has a two directional structure with horizontal and vertical axes showing the time and frequency domain respectively. A slot is the smallest unit of resource in a frame, which occupies space both in time and frequency domain. A burst is a set of slots using the same MCS. The total number of slots in the frame depends on the subcarrier permutation method. For numerical applications, we focus on PUSC, though our model is valid for any permutation scheme. In fact, a slot always carries 48 subcarriers whatever the type of subcarrier permutation used. In the DL sub-frame, the first part contains Preamble, Frame Control Header (FCH), UL_MAP and DL_MAP. Preamble is used for synchronization. FCH provides length and encoding of two MAP messages and information about usable subchannels. The data mapping for users resides in the MAP messages. Their size depends on the number of scheduled users in the frame.

One of the important features of IEEE 802.16e is link adaptation: different 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 choice of the right MCS is done according to the 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.

The scheduling algorithm is responsible for allocating radio resources of a frame (or of a group of frames) to active users. In wireless networks, scheduling may take into account their radio link quality. In this work, we have considered three traditional schemes. The slot

fairness scheduling allocates the same number of slots to all active users. The throughput fairness scheduling ensures that all active users have the same instantaneous throughput. The opportunistic scheduling gives all resources to active users with the best channel.

Let us now define the notations concerning the WiMAX system needed in this chapter:

- N_S is the total number of slots available for data transmission in the downlink part of the TDD frame. As mentioned before, N_S depends on the system bandwidth, the frame duration, the DL/UL ratio, the permutation scheme and the overhead.
- T_F is the TDD frame duration: $T_F = 5$ ms.
- Radio channel states are denoted MCS_k, $1 \leq k \leq K$, where K is the number of MCS. By extension, we denote MCS_0 the outage state.
- m_k is the number of bits transmitted per slot by an MS using MCS_k. Recall that the number of bits transmitted per slot is independent of the permutation method and is thus constant for a given MCS. For the particular case of outage, $m_0 = 0$.

1.5 WiMAX Analytical modeling

This section provides the development of our generic analytical model for WiMAX networks. We consider a single WiMAX cell handling the data traffic. This study targets the analysis of bottleneck, i.e., the radio link, and focuses on the downlink part, which is assumed to be a critical resource in asymmetric data traffic.

1.5.1 System modeling

The development of our analytical model is based on several assumptions related either to the system or the traffic. All of them will be discussed in Section 1.5.4, and, as developed in that section, most of them can be relaxed, if necessary, by slightly modifying the basic model.

System assumptions

1. The size of the DL_MAP and UL_MAP parts of the TDD frame is assumed to be constant and independent of the number of concurrent active mobiles. As a consequence, the total number of slots available for data transmission in the downlink part is constant and equals N_S.

2. We assume that the number of simultaneous mobiles that can be multiplexed in one TDD frame is not limited. As a consequence, any connection demand will be accepted and no blocking can occur.

3. At any given time, if there is only one active user, we assume that the scheduler can allocate all the available slots for its transfer.

Channel assumptions

4. The coding scheme used by a given mobile can change very often because of the high variability of the radio link quality. We assume that each mobile sends feedback channel estimation on a frame by frame basis, and thus, the base station can change its coding scheme every frame. Since we do not make any distinction between users and consider all mobiles 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 mobile has a probability p_k to use MCS_k (including outage).

Traffic assumptions

5. All the users have the same traffic characteristics. In addition, we don't consider any QoS differentiation.

6. We don't take handover into account.

7. We assume that there is a fixed number N of mobile stations (MS) that are sharing the available bandwidth of the cell.

8. Each of the N mobiles 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 the embedded objects). As the downloading duration depends on the system load and the radio link quality, 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. Unlike ON periods, OFF periods must then be characterized by their duration.

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

1.5.2 Analytical model

In order to develop our WiMAX analytical model we first consider a system with a single coding scheme (i.e., $K = 1$) and no outage. We denote the number of bits transferred by any slot by m $(= m_1)$, and define μ, the average departure rate, as:

$$\mu = \frac{m\, N_S}{\bar{x}_{on}\, T_F}. \tag{1.1}$$

We also define λ, the inverse of the average reading time, as:

$$\lambda = \frac{1}{\bar{t}_{off}}. \tag{1.2}$$

With all the assumptions presented in the previous subsection, this basic system can be modeled by a simple Continuous Time Markov Chain (CTMC) made up of $N + 1$ states. A state n of this chain $(0 \leq n \leq N)$ corresponds to the total number of concurrent active mobiles, i.e., mobiles that are in ON period.

- A transition out of a generic state n to state $n+1$ occurs when a mobile in OFF period starts its transfer. This "arrival" transition is performed with a rate $(N - n)\lambda$. It corresponds to one mobile among the $(N - n)$ in OFF period, ending its reading.

- A transition out of a generic state n to state $n-1$ occurs when a mobile in ON period completes its transfer. This "departure" transition is always performed with a rate μ corresponding to the total departure rate of the frame.

It turns out that this basic but unrealistic model is equivalent to the classical Engset model [9].

We now go back to the real system including several MCS ($K > 1$). Because of coding scheme diversity, the average departure rate is no longer constant. It actually depends both on the active mobile population and on the scheduling policy integrated into the system. The analytical model we propose keeps the same birth-and-death structure but integrates departure rates $\mu(n)$ that depend on the current state n as shown in Fig. 1.2.

The main difficulty consists in estimating accurately the average departure rates $\mu(n)$ of this model. In order to do so, we first express $\mu(n)$ as follows:

$$\mu(n) = \frac{\bar{m}(n)\, N_S}{\bar{x}_{on}\, T_F}, \tag{1.3}$$

where $\bar{m}(n)$ is the average bit rate 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. Also, $\bar{m}(n)$ strongly depends on n, because the average bit rate per slot must be estimated by considering all possible distributions of the n mobiles between the $K+1$ possible coding schemes (including outage). Finally, the average bit rates $\bar{m}(n)$ also depend on the scheduling policy. More precisely, for each possible mobiles distribution, the scheduling policy defines the quantity of slots given to each of the n mobiles that corresponds to the coding scheme they use.

At this step, our analytical model can represent any WiMAX system provided the average bit rates $\bar{m}(n)$ can be estimated. In Section 1.5.3 we develop a generic analytical expression of these rates, whereas in Section 1.6 we present their detailed expressions depending on three specific scheduling policies.

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. 1.2):

$$\pi(n) = \frac{N!}{(N-n)!}\left(\frac{T_F}{N_S}\right)^n \frac{\rho^n}{\prod_{i=1}^{n}\bar{m}(n)}\pi(0) \qquad \text{, with } \rho = \frac{\bar{x}_{on}}{\bar{t}_{off}}, \tag{1.4}$$

and $\pi(0)$ is obtained by normalization.

The performance parameters of this system can be derived from the steady-state probabilities as follows. The average utilization $\bar{U}$ of the TDD frame is given by:

$$\bar{U} = \sum_{n=1}^{N} \pi(n) \min\left(n\frac{\bar{x}_{on}}{N_S\,\bar{m}(n)}, 1\right). \tag{1.5}$$

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

$$\bar{Q} = \sum_{n=1}^{N} n\,\pi(n). \tag{1.6}$$

X_d, the mean number of departures (mobiles completing their transfer) per unit of time, is obtained as:

$$\bar{X}_d = \sum_{n=1}^{N} \pi(n)\,\mu(n). \tag{1.7}$$

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

$$\bar{t}_{on} = \frac{\bar{Q}}{\bar{X}_d}. \tag{1.8}$$

We finally compute the average throughput $\bar{X}$ obtained by each mobile in active transfer as:

$$\bar{X} = \frac{\bar{x}_{on}}{\bar{t}_{on}}. \tag{1.9}$$

1.5.3 Generic average bit rates

We now develop generic expressions of the average bit rates $\bar{m}(n)$ without and with outage.

Without outage

We first consider a system without outage. In order to illustrate the derivation of the average bit rates per slot, we first consider a situation with 2 active mobiles (denoted as MS1 and MS2) in a system with 2 MCS ($K = 2$), and develop the expression of $\bar{m}(2)$. MCS_1 is used with a probability p_1 and allows to transfer m_1 bits per slot. MCS_2 is used with a probability p_2 and allows to transfer m_2 bits per slot. We denote the average bit rate per slot in the TDD frame for one particular configuration having j_1 mobiles using MCS_1 and j_2 mobiles using MCS_2 ($j_1 + j_2 = 2$) by $\bar{m}(j_1, j_2)$. There are 3 possible configurations:

- $MCS_1 = 2$ MS and $MCS_2 = 0$ MS. This configuration occurs with a probability $p_1\, p_1$. Whatever the scheduling policy, the corresponding average bit rate $\bar{m}(2, 0)$ is obviously given as:

$$\bar{m}(2, 0) = m_1; \tag{1.10}$$

- $MCS_1 = 0$ MS and $MCS_2 = 2$ MS. Similarly, with a probability $p_2\, p_2$, we have:

$$\bar{m}(0, 2) = m_2; \tag{1.11}$$

- $MCS_1 = 1$ MS and $MCS_2 = 1$ MS. This configuration can correspond to 2 different distributions of the 2 mobiles: MS1 = MCS_1 and MS2 = MCS_2, or MS1 = MCS_2 and MS2 = MCS_1. The associated probability is $2\, p_1\, p_2$, as both distributions have equal probabilities. The corresponding average bit rate $\bar{m}(1, 1)$ can be expressed as:

$$\bar{m}(1, 1) = m_1 x_1(1, 1) + m_2 x_2(1, 1), \tag{1.12}$$

  where $x_k(1, 1)$ is the proportion of the resource that is associated to mobiles using MCS_k, which strongly depends on the scheduling policy.

We finally express the average bit rate per slot when there are 2 active mobiles in the system as:

$$\bar{m}(2) = \sum_{j_1=0}^{2} \bar{m}(j_1, 2-j_1) \binom{2}{j_1} p_1^{j_1} p_2^{2-j_1}, \tag{1.13}$$

where $\binom{2}{j_1}$ is a binomial coefficient that gives the number of distributions corresponding to the configuration of j_1 mobiles using MCS_1 and $2-j_1$ mobiles using MCS_2. As a generalization, one can convince himself easily that the average bit rate per slot, $\bar{m}(n)$, when there are n active users, can be expressed as follows:

$$\begin{aligned} \bar{m}(n) &= \sum_{\substack{(j_1,...,j_K)=(0,...,0)| \\ j_1+...+j_K=n}}^{(n,...,n)} \bar{m}(j_1,...,j_K) \binom{n}{j_1,...,j_K} \left(\prod_{k=1}^{K} p_k^{j_k} \right) \\ &= \sum_{\substack{(j_1,...,j_K)=(0,...,0)| \\ j_1+...+j_K=n}}^{(n,...,n)} \sum_{k=1}^{K} m_k j_k x_k(j_1,...,j_K) \binom{n}{j_1,...,j_K} \left(\prod_{k=1}^{K} p_k^{j_k} \right), \end{aligned} \tag{1.14}$$

where $\binom{n}{j_1,...,j_K}$ is the multinomial coefficient and $x_k(j_1,...,j_K)$ is the proportion of resource given to MS using MCS_k, when the current distribution of the n mobiles among the K coding schemes is $(j_1,...,j_K)$. Let us emphasize that this expression has a $\mathcal{O}(n^K)$ complexity, where K, the number of different coding schemes, is usually low. Section 1.6.4 will show that this complexity can be drastically reduced without any significant impact on the accuracy of $\bar{m}(n)$ values.

With outage

We now come back to a system with possible outage (MCS_0 used with a probability p_0). Relation (1.14) can be extended straightforwardly as the j_0 mobiles in outage of a given distribution do not contribute to the sharing of resource:

$$\bar{m}(n) = \sum_{\substack{(j_0,j_1,...,j_K)=(0,0,...,0)| \\ j_0+j_1+...+j_K=n \\ j_0 \neq n}}^{(n,n,...,n)} \left(\sum_{k=1}^{K} m_k j_k x_k(j_1,...,j_K) \right) \binom{n}{j_0,j_1,...,j_K} \left(\prod_{k=0}^{K} p_k^{j_k} \right). \tag{1.15}$$

1.5.4 Discussion of the modeling assumptions

Our Markovian model is based on several system and traffic assumptions presented in Section 1.5.1. We now discuss these assumptions one by one (item numbers are related to the corresponding assumptions), evaluate their accuracy, and provide, if necessary and possible, extensions and generalization propositions.

1. As described in Section 1.4, DL_MAP and UL_MAP are located in the downlink part of the TDD frame. They contain the information elements that allow mobiles to identify the slots to be used. The size of these MAPs, and as a consequence the number N_S of available slots for downlink data transmissions, depends on the number of mobiles scheduled in the TDD frame. In order to relax assumption 1, we can express the number of data slots, $N_S(n)$, as a function of n, the number of active users. This dependency can be easily integrated in the model by replacing N_S^n by $\prod_{i=1}^{n} N_S(n)$ in relation (1.4), and N_S by $N_S(n)$ in relations (1.3) and (1.5).

2. A limit n_{max} on the total number of mobiles that can simultaneously be multiplexed on the TDD frame, can easily be introduced in the model, if required. The corresponding Markov chain shown in Fig. 1.2, indeed, has just to be truncated to this limiting state (i.e., the last state becomes $\min(n_{max}, N)$). As a result, a blocking can now occur when a new transfer demand arrives and the limit is reached. The blocking probability can easily be derived from the Markov chain [5].

3. In some cellular networks (e.g. (E)GPRS), mobile stations have limited transmission capabilities because of hardware considerations. This constraint defines the maximum throughput the network interface can reach or the maximum number of resource units that can be used by the mobiles. Such limitations add a slight complexity to the development of model, as one single mobile may not be able to use all the available slots. This characteristic has been introduced in the case of (E)GPRS networks [5, 22] and can be applied to WiMAX networks by simply modifying the departure rates of the first states of the Markov chain (i.e., replacing N_S by $\min(n\,d, N_S)$ in relation (1.3), where d is the maximum number of slots a mobile can use in downlink).

4. Radio channel may be highly variable (i.e., conditions change from one frame to another)

or it may vary with some memory (i.e., conditions are maintained during a number of frames). Our analytical model only depends upon stationary probabilities of different coding schemes whatever be the radio channel dynamics. This approach is authenticated through simulations in Section 1.7.

5. All mobiles in the considered system have statistically the same traffic characteristics. More complex systems with multiple-traffic and/or differentiation between users would naturally result into more complex models that are not addressed here.

6. As our main concern is dimensioning, we do not take handover into account and consider the fixed mobile population in a stationary manner. However, mobility effects are taken into account in the channel model by means of radio conditions variation.

7. Poisson processes are currently used in the case of a large population of users, assuming independence between the arrivals and the current population of the system. As we focus on the performance of a single cell system, the potential population of users is relatively small. The higher the number of on-going data connections, the less likely the arrival of new ones. Poisson processes are thus a non-relevant choice for our models. In addition, the finite population assumption is used typically for network planning when geo-marketing data allows the prediction of the active mobile population that will be served by the cell (for a network in service, traffic statistics can also provide estimates of this population). Note, however, if the Poisson assumption has to be made for connection demand arrivals, one can directly modify the arrival rates of the Markov chain (i.e., replace the state-dependent rates $(N - n)\lambda$ by some constant value, and limit the number of states of the Markov chain as explained above in point 2).

8. Each mobile is supposed to generate infinite length ON/OFF session traffic. In the context of (E)GPRS networks [4, 6], we have studied an extension to finite length sessions, where each mobile generates ON/OFF traffic during a session and does not generate any traffic during an inter-session. We show in these studies that a very simple transformation of traffic characteristics, that increases OFF periods by a portion of the inter-session period, enables to derive the average performance from the infinite length session model. The accuracy of this transformation is related to the insensibility of the average performance parameters with regards to the traffic distributions (see next

point). An equivalent transformation can be applied to our WiMAX model, even if it is no longer processor sharing. Until any theoretical result can be proven, the resulting transformation remains a good approximation.

9. Memoryless traffic distributions are strong assumptions that are validated by several theoretical results. Several studies on insensitivity (see, e.g., [7, 11, 14]) have shown (for processor sharing systems) that the average performance parameters are insensitive to the distribution of ON and OFF periods. As here, we are not able to formally demonstrate that this result also holds for our WiMAX model, we present in Section 1.7 a comparison of the system performance obtained by simulation for several traffic distributions (exponential and Pareto), and our analytical model. These results tend to prove that insensibility still holds or is at least a good approximation. Thus, memoryless distributions are the most convenient choices to model traffic.

1.6 Scheduling policy modeling

We now present the analytical model adaptation to different scheduling policies. For each of them we provide closed-form expressions for the average bit rates per slot, $\bar{m}(n)$.

1.6.1 Slot sharing fairness

We study a scheduling policy providing fairness in slot sharing. Each time-step, the scheduler equally shares the N_S slots among the active users that are not in outage.

Slot sharing fairness without outage

First we don't take into account outage (only coding schemes MCS_k, $1 \leq k \leq K$, are used). If at a given time-step, there are n active mobiles, then each of them receives a portion N_S/n of the whole resource. As a consequence, the proportion of the resource that is associated to mobiles using MCS_k, is constant for any k and for any possible distribution $(j_1, ..., j_K)$

of the n mobiles among the K coding schemes, and is thus given by:

$$x_k(j_1, ..., j_K) = \frac{1}{n}. \tag{1.16}$$

By replacing these equal proportions in generic expression (1.14), the average bit rate per slot, $\bar{m}(n)$, when there are n active users, becomes:

$$\bar{m}(n) = \sum_{\substack{(j_1,...,j_K)=(0,...,0)| \\ j_1+...+j_K=n}}^{(n,...,n)} \left(\sum_{k=1}^{K} \frac{m_k j_k}{n} \right) \binom{n}{j_1, ..., j_K} \left(\prod_{k=1}^{K} p_k^{j_k} \right). \tag{1.17}$$

After a few simplifications, we obtain:

$$\bar{m}(n) = \sum_{\substack{(j_1,...,j_K)=(0,...,0)| \\ j_1+...+j_K=n}}^{(n,...,n)} \sum_{\substack{k=1| \\ j_k \neq 0}}^{K} m_k p_k \binom{n-1}{j_1, ..., j_{k-1}, ..., j_K} \left(p_1^{j_1} ... p_k^{j_k - 1} ... p_K^{j_K} \right). \tag{1.18}$$

By rearranging carefully the terms of the summations, we can show that this expression can drastically be simplified as:

$$\bar{m}(n) = \sum_{k=1}^{K} m_k p_k = \bar{m} \tag{1.19}$$

This nice and very simple expression shows us that, when there is no outage, the average bit rates $\bar{m}(n)$ associated with the slot sharing fairness policy are constant, and can be simply seen as an average bit rate $\bar{m}$.

Slot sharing fairness with outage

If we now consider outage (MCS_0 is used with a probability p_0), the expression of the average bit rates per slot becomes:

$$\begin{aligned}\bar{m}(n) &= \sum_{\substack{(j_0,j_1,...,j_K)=(0,0,...,0)|\\ j_0+j_1+...+j_K=n\\ j_0\neq n}}^{(n,n,...,n)} \left(\sum_{k=1}^{K}\frac{m_k j_k}{n-j_0}\right)\binom{n}{j_0,j_1,...,j_K}\left(\prod_{k=0}^{K}p_k^{j_k}\right)\\ &= \sum_{\substack{(j_0,j_1,...,j_K)=(0,0,...,0)|\\ j_0+j_1+...+j_K=n\\ j_0\neq n}}^{(n,n,...,n)} \frac{n!}{n-j_0}\left(\sum_{k=1}^{K}m_k j_k\right)\left(\prod_{k=0}^{K}\frac{p_k^{j_k}}{j_k!}\right), \end{aligned} \tag{1.20}$$

as the resource has now to be shared only among the $n - j_0$ users that are not in outage. It is important to note that the expression of the average bit rates cannot be further simplified, and is no longer constant (see Section 1.6.4).

1.6.2 Instantaneous throughput fairness

We now consider a scheduling policy that shares the resource in order to provide the same instantaneous throughput to all active users. Hence, at a given time-step, mobiles using MCS with a low bit rate per slot will obtain proportionally more slots than the mobiles using a MCS with a high bit rate per slot.

Throughput fairness without outage

Let us first consider that there is no possible outage. Recall that $x_k(j_1, ..., j_K)$ is the proportion of the resource that is associated by the scheduler to mobiles using coding scheme MCS_k, when the current distribution of the mobiles is $(j_1, ..., j_K)$. In order to respect instantaneous

throughput fairness between active users, the $x_k(j_1, ..., j_K)$ must be such that:

$$m_k x_k(j_1, ..., j_K) = C \text{ for any } k, \tag{1.21}$$

where C is a constant such that $\sum_{k=1}^{K} j_k x_k(j_1, ..., j_K) = 1$, thus:

$$C = \frac{1}{\sum_{k=1}^{K} \frac{j_k}{m_k}}. \tag{1.22}$$

By replacing the proportions $x_k(j_1, ..., j_K)$ in generic expression (1.14), the average bit rate per slot, $\bar{m}(n)$, when there are n active users, becomes:

$$\bar{m}(n) = \sum_{\substack{(j_1,...,j_K)=(0,...,0)| \\ j_1+...+j_K=n}}^{(n,...,n)} \frac{n}{\sum_{k=1}^{K} \frac{j_k}{m_k}} \binom{n}{j_1, ..., j_K} \left(\prod_{k=1}^{K} p_k^{j_k} \right) = \sum_{\substack{(j_1,...,j_K)=(0,...,0)| \\ j_1+...+j_K=n}}^{(n,...,n)} \frac{n\, n! \prod_{k=1}^{K} \frac{p_k^{j_k}}{j_k!}}{\sum_{k=1}^{K} \frac{j_k}{m_k}}. \tag{1.23}$$

Throughput fairness with outage

If we now consider outage, the expression of the average bit rates per slot simply becomes:

$$\bar{m}(n) = \sum_{\substack{(j_0,j_1,...,j_K)=(0,0,...,0)| \\ j_0+j_1+...+j_K=n \\ j_0 \neq n}}^{(n,n,...,n)} \frac{(n-j_0)\, n! \prod_{k=0}^{K} \frac{p_k^{j_k}}{j_k!}}{\sum_{k=1}^{K} \frac{j_k}{m_k}}. \tag{1.24}$$

1.6.3 Opportunistic scheduling

We finally study an opportunistic scheduling policy where all the resources are given to users having the highest transmission bit rate, i.e. the better radio conditions and hence the better MCS. Without loss of generality, we assume, in this section, that the coding schemes are

classified in increasing order: $m_1 < m_2 < ... < m_K$. And even if it is still possible to derive the average bit rates from generic expressions (1.14) and (1.15) (without and with outage), we prefer to give here a more intuitive and strictly equivalent derivation.

Opportunistic without outage

We consider a system with n current active mobiles. We denote by $\alpha_i(n)$ the probability of having at least one active user (among n) using MCS_i and none using a MCS giving higher transmission rates (i.e. MCS_j with $j > i$). As a matter of fact, $\alpha_i(n)$ corresponds to the probability that the scheduler gives at a given time-step all the resource to mobiles that use MCS_i. As a consequence, we can express the average bit rate per slot when there are n active users as:

$$\bar{m}(n) = \sum_{i=1}^{K} \alpha_i(n) m_i. \tag{1.25}$$

In order to calculate the $\alpha_i(n)$, we first express $p_{\leq i}(n)$, the probability that there is no mobile using a MCS higher than MCS_i:

$$p_{\leq i}(n) = \left(1 - \sum_{j=i+1}^{K} p_j\right)^n. \tag{1.26}$$

Then, we calculate $p_{=i}(n)$, the probability that there is at least one mobile using MCS_i conditioned by the fact that there is no mobile using a better MCS:

$$p_{=i}(n) = 1 - \left(1 - \frac{p_i}{\sum_{j=1}^{i} p_j}\right)^n. \tag{1.27}$$

$\alpha_i(n)$ can thus be expressed as:

$$\alpha_i(n) = p_{=i}(n)\, p_{\leq i}(n). \tag{1.28}$$

Opportunistic with outage

It is easy to show that the previous development remains the same when some mobiles can be in outage. Indeed, as soon as there is at least one mobile using a real coding scheme MCS_k ($k \neq 0$), the scheduler gives no resource to mobiles using a "lower" MCS (including mobiles in outage). As a result, the average bit rates per slot have the same expression (with a single modification in relation (1.27) where index j of the sum must vary from 0 to i).

1.6.4 Analytical asymptotic study

As a side study in the modeling of the average bit rates $\bar{m}(n)$, we can observe asymptotic behavior of $\bar{m}(n)$ functions. Fig. 1.3 shows the evolution of $\bar{m}(n)$ when n increases for the three studied scheduling policies, in the general case where some mobiles can be in outage. We can notice that the three resulting functions $\bar{m}(n)$ rapidly tend to an asymptote, as the number of active users n increases. We thus derive in the following subsections the analytical expressions of these asymptotes for each scheduling policy. Note that one can benefit from this quick asymptotical behavior to avoid the calculation of the $\bar{m}(n)$ for large values of n (by replacing, after a threshold, the exact value by the corresponding asymptote value).

Slot fairness asymptote

In the case of slot fairness scheduling, as the number of active users grows, the proportion of mobiles using MCS_k tends to p_k. If we denote the number of such mobiles by J_k, when $n \to \infty$, we have $J_k \sim p_k\, n$. As the resources are equally shared among $n - J_0$ mobiles that are not in outage, the limiting value of the average bit rates is given by:

$$m_S(\infty) = \lim_{n\to\infty} \bar{m}(n) = \lim_{n\to\infty} \sum_{k=1}^{K} m_k \frac{J_k}{n - J_0} = \frac{\sum_{k=1}^{K} m_k p_k}{1 - p_0}. \tag{1.29}$$

Throughput fairness asymptote

We now detail the asymptote corresponding to the instantaneous throughput fairness policy. Again, the number of mobiles using MCS_k, when $n \to \infty$, is $J_k \sim p_k\, n$. Every such mobile obtains a proportion x_k of the resource such that $\sum_{k=1}^{K} J_k x_k = 1$. In order to respect the fairness of the scheduling policy, these proportions must satisfy the following relation:

$$m_k x_k = C = \frac{1}{\sum_{k=1}^{K} \frac{J_k}{m_k}} \text{ for any } k \neq 0. \tag{1.30}$$

Note that mobiles in outage do not use any resource (and thus, $x_0 = 0$). By combining these relations, we obtain the expression of the asymptote value:

$$m_X(\infty) = \lim_{n\to\infty} \bar{m}(n) = \lim_{n\to\infty} \sum_{k=1}^{k} m_k J_k x_k = \frac{1 - p_0}{\sum_{k=1}^{K} \frac{p_k}{m_k}}. \tag{1.31}$$

Opportunistic scheduling asymptote

The asymptote value of $\bar{m}(n)$ for opportunistic scheduling simply corresponds to the highest bit rate per slot (obtained with the best coding scheme). Actually, as the number of active users grows, the probability of having at least one mobile using the best MCS tends to 1. Thus, we have:

$$m_{opp}(\infty) = \lim_{n\to\infty} \bar{m}(n) = m_K. \tag{1.32}$$

1.7 Validation

In this section, we discuss the validation of our analytical model through extensive simulations. We also show its robustness when traffic and channel models are complexified. For this purpose, a simulator has been developed that implements an ON/OFF traffic generator and

a wireless channel for each user, and a centralized scheduler that allocates radio resources, i.e., slots, to active users on a frame by frame basis. In a first phase, we validate our analytical model through simulations. In this *validation study*, the modeling assumptions (related to scheduling, traffic and channel models) are reproduced in the simulator. The assumptions are related to scheduling, traffic and channel models. This phase shows that describing the state of the system by the aggregation of all active users (whatever the distribution of their coding schemes) is a very good modeling approximation. It also validates the analytical expression of the average bit rates $\bar{m}(n)$. In a second phase, the *robustness study*, we relax the assumptions made for the analytical model by considering more realistic models for traffic and radio channel variations. By carrying out a comparison with simulation results, we thus show how robust the analytical model reacts towards these relaxations.

1.7.1 Simulation Models

We now detail the simulation models before presenting the simulations results for the validation and robustness studies.

System Parameters

As in the previous sections, we consider a single WiMAX cell and study the downlink. Radio resources are thus made of time-frequency slots in the downlink TDD sub-frame. The number of slots depends on the system bandwidth, the frame duration, the downlink/uplink ratio, the subcarrier permutation (PUSC, FUSC, AMC), and the protocol overhead (preamble, FCH, maps). System bandwidth is assumed to be 10 MHz. The duration of one TDD frame of WiMAX is 5 ms and the downlink/uplink ratio is considered to be 2/3. Although a slot is made of the same number of data subcarriers whatever the subcarrier permutation is, their total number varies. For the purpose of simulation, PUSC has been kept as a reference. We assume for the sake of simplicity that the protocol overhead is of fixed length (2 symbols) although in reality it is a function of the number of scheduled users. These parameters lead to a number of data slots (excluding overhead) per TDD downlink sub-frame of $N_S = 450$.

Traffic Parameters

In our analytical model, we consider an elastic ON/OFF traffic. Mean values of ON data volume (main page and embedded objects) and OFF period (reading time), are 3 Mbits and 3 s respectively.

In the first phase (validation study), we assume that the ON data volume is exponentially distributed, as it is the case in the analytical model assumptions. Although well adapted to Markov theory based analysis, exponential law does not always fit the reality for data traffic. This is the reason why we consider truncated Pareto distributions in the second phase (the robustness study). Recall that the mean value of the truncated Pareto distribution is given by:

$$\bar{x}_{on} = \frac{\alpha b}{\alpha - 1}\left[1 - (b/q)^{\alpha-1}\right], \tag{1.33}$$

where α is the shape parameter, b is the minimum value of Pareto variable and q is the cutoff value for truncated Pareto distribution. Two values of q are considered: low and high. These have been taken as hundred times and thousand times the mean value respectively. The mean value in both cases (high and low cutoff) is 3 Mbits for the sake of comparison with the exponential model. The value of $\alpha = 1.2$ has been adopted from [10]. The corresponding values of parameter b for high and low cutoff are calculated using relation (1.33). Traffic parameters are summarized in Tab. 1.1.

Parameter	**Value**
Number of MS in the system N	up to 50
Mean ON data volume $\bar{x}_{on}$	3 Mbits
Mean OFF duration $\bar{t}_{off}$	3 s
Pareto parameter α	1.2
Pareto low cutoff q	300 Mbits
Pareto high cutoff q	3000 Mbits
Pareto parameter b for low cutoff	712926 bits
Pareto parameter b for high cutoff	611822 bits

Table 1.1: Traffic parameters.

Channel Models

The number of bits per slot an MS is likely to receive depends on the chosen MCS, which in turn depends on its radio channel conditions. The choice of a MCS is based on SINR measurements and SINR thresholds. A generic method for describing the channel between the BS (Base Station) and a MS is to model the transitions between MCS by a finite state Markov chain (FSMC). The chain is discrete time and transitions occurs every L frames, with $L\,T_F < \bar{t}_{coh}$, where $\bar{t}_{coh}$ is the coherence time of the channel. In our case, for the sake of simplicity, $L = 1$. Such a FSMC is fully characterized by its transition matrix $P_T = (p_{ij})_{0 \leq i,j \leq K}$. Note that an additional state (state 0) is introduced to take into account outage (when SINR is below the minimum radio quality threshold). Stationary probabilities p_k provide the long-term probabilities for a MS to receive data with MCS k.

In our analytical study, channel model is assumed to be memoryless, i.e., MCS are independently drawn from frame to frame for each user, and the discrete distribution is given by $(p_i)_{0 \leq i,j \leq K}$. This corresponds to the case where $p_{ij} = p_j$ for all i. This simple approach, referred as the *memoryless channel model*, is considered in the validation study, which exactly reproduces the assumptions of the analysis. Let $P_T(0)$ be the transition matrix associated to the memoryless model.

In the robustness study, we introduce two additional channel models with memory. In these models, the MCS observed for a given MS in a frame depends on the MCS observed in the previous frame according to the FSMC presented above. The transition matrix is derived from the following equation:

$$P_T(a) = aI + (1-a)P_T(0) \quad 0 \leq a \leq 1,$$

where I is the identity matrix and parameter a is a measure of the channel memory. A mobile actually maintains its MCS for a certain duration with mean $\bar{t}_{coh} = 1/(1-a)$. With $a = 0$, the transition process becomes memoryless. On the other extreme, with $a = 1$, the transition process will have infinite memory and mobiles will never change their MCS. For simulations, we have taken a equal to be 0.5, so that the channel is constant in average 2

frames. This value is consistent with the coherence time given in [23] for a 45 Km/h speed mobility in a 2.5 GHz bandwidth system. We call the case where all MS have the same channel model with memory ($a = 0.5$), the *average channel model*. Note that the stationary probabilities of the average channel model are the same as those of the memoryless model.

As the channel depends on the BS-MS link, it is possible to refine the previous approach by considering part of the MS to be in a "bad" state, and the rest in a "good" state. Bad and good states are characterized by different stationary probabilities but have the same coherence time. In the so called *combined channel model*, half of the MS are in a good state, the rest in a bad state, and a is kept to 0.5 for both populations. For the sake of comparison, the overall MCS probabilities in the combined model are the same as those of the memoryless and average models. Three models are thus considered: the memoryless, the average, and the combined channel models. In Tab. 1.2, considered MCS (including outage) are given, and for each of them, the numbers of bits transmitted per slot are also listed.

Channel state $\{0, ..., K\}$	**MCS and outage**	**Bits per slot** m_k
0	Outage	$m_0 = 0$
1	QPSK-1/2	$m_1 = 48$
2	QPSK-3/4	$m_2 = 72$
3	16QAM-1/2	$m_3 = 96$
4	16QAM-3/4	$m_4 = 144$

Table 1.2: Channel parameters.

Channel stationary probabilities are given in Tab. 1.3. The respective MCS stationary probabilities for good and bad channel types can be obtained for example by performing system level Monte Carlo simulations and recording channel statistics close (good state) or far (bad state) from the BS. Stationary probabilities for the combined model are obtained by averaging corresponding values of good and bad model stationary probabilities.

Channel model	**Memoryless**	**Average**	**Combined**	
			good 50% MS	**bad** 50% MS
a	0	0.5	0.5	0.5
p_0	0.225	0.225	0.020	0.430
p_1	0.110	0.110	0.040	0.180
p_2	0.070	0.070	0.050	0.090
p_3	0.125	0.125	0.140	0.110
p_4	0.470	0.470	0.750	0.190

Table 1.3: Stationary probabilities for three channel models

Scheduling simulation

The simulator implements the three scheduling schemes considered in this study, i.e., opportunist, fair in throughput and fair in slots. On a frame-by-frame basis, the scheduler allocates the downlink slots to the active users according to their radio conditions (their MCS) and the scheduling policy. As already mentioned, the scheduler does not allocate resources to active users in outage. The computation of number of slots to be allocated to each user is detailed hereafter. In a given frame, the number of slots allocated to active users should satisfy the following condition:

$$N_S = \sum_{k=0}^{K} N_S^{(k)} n^{(k)},$$

where $N_S^{(k)}$ is the number of slots allocated by the scheduler to a MS using MCS k and $n^{(k)}$ is the number of active mobiles using MCS k. Note that $N_S^{(k)}$ depends on the scheduling scheme and that the number of active users verifies $n = \sum_{k=0}^{K} n^{(k)}$. The way $N_S^{(k)}$ is chosen by the scheduler is detailed below in the scheduling pseudo-code.

Scheduling pseudo-code

Let $\mathbf{K_F} \subset [0...K]$ be the set of MCS used by active MS in the considered frame.

▷ Opportunist

find $k_{max} = max(\mathbf{K_F})$

$N_S^{(k)} = \frac{N_S}{n^{(k_{max})}}$ for $k = k_{max}$

$N_S^{(k)} = 0$ for all $k \neq k_{max}$

▷ Fairness in slot

$N_S^{(k)} = 0$ for $k = 0$

$N_S^{(k)} = \frac{N_S}{\sum_{k=0}^{K} n^{(k)}}$ for $k \neq 0$

▷ Fairness in throughput

if $(k = 0)$ **then** $N_S^{(k)} = 0$

else $N_S^{(k)} = \frac{N_S/m_k}{\sum_{k=1}^{K} \frac{n^{(k)}}{m_k}}$

The value of $N_S^{(k)}$ determined by the scheduling process may or may not be an integer. In case it is not, it is rounded down to the closest integer. It results into some spare number of slots that are allocated to the active users (not in outage) in a Round Robin fashion.

1.7.2 Simulation Results

In this section, we first present a comparison between the results obtained through our analytical model and scheduling simulator. The output parameters in consideration are $\bar{U}$, $\bar{X}$, and $\pi(n)$ (see Section 1.5.2).

Validation Study

In this study, simulations take into account the same traffic and channel assumptions as those of the analytical model. However, in simulator MCS of users are determined on per frame basis and scheduling is carried out in real time, based on MCS at that instant. The analytical model on the other hand, considers stationary probabilities of MCS only. Distributions of ON data volume and OFF period are exponential and the memoryless channel model is considered.

Fig. 1.4(a, b) show respectively the average channel utilization ($\bar{U}$) and the average instantaneous throughput per user ($\bar{X}$) for the three scheduling schemes. It is clear that simulation and analytical results show a good agreement: for both utilization and throughput, the maximum relative error stays below 6% and the average relative error is less than 1%. Note that the analytical results have been obtained instantaneously whereas simulations have run for several days.

Fig. 1.4(c) further proves that our analytical model is a very good description of the system: stationary probabilities $\pi(n)$ obtained by either simulations or analysis are compared for a given total number $N = 50$ of MS. Again, results show a perfect match with an average relative error always below 9%. This means that not only average values of the output parameters can be derived from stationary probabilities with a high accuracy but also higher moments.

At last, Fig. 1.5 shows the validation for three different loads (1, 3 and 5 Mbps). Our model shows a comparable accuracy for all three load conditions with a maximum relative error of about 5%.

Robustness Study

We now move to the robustness study, where assumptions concerning traffic and channel models made by the analysis are relaxed in simulations. Note that we have run extensive simulations corresponding to various traffic and channel models, that give very similar results.

In order to check the robustness of analytical model towards distribution of ON data volumes, simulations are carried out for exponential and truncated pareto (with low and high cutoff). The results for this analysis are shown in Fig. 1.6. The average relative error between analytical results and simulations stays below 10% for all sets. It is clear that considering a truncated Pareto distribution has little influence on the design parameters. This is mainly due to the fact that the distribution is truncated and is thus not heavy tailed. But even with a high cutoff value, the exponential distribution provides a very good approximation.

Until now we have always considered the memoryless channel model. In order to check the robustness of our analytical model with respect to the channel memory, we now compare the analytical results with simulation for the three pre-cited channel models: memoryless, average and combined (with stationary probabilities given in Tab. 1.3). If we look at the plot of Fig. 1.7, we can say that even for a complex channel, our analytical model shows considerable robustness with an average relative error below 7%. We can thus deduce that for designing a WiMAX network, channel information is almost completely included in the stationary probabilities of the MCS.

1.8 Performance analysis

In this section, we use the analytical models developed in Sections 1.5 and 1.6 to derive performance curves. We give a first set of general conclusions regarding system behaviors. Note however that this study does not intend to decide which scheduling policy is to be selected, as this would require a wider analysis. We consider a TDD frame with $N_S = 450$ downlink data slots. The system contains $N = 50$ mobiles with the following traffic characteristics: the average size of a downloaded elements is $\bar{x}_{on} = 1$ Mbits and the average reading time is $\bar{t}_{off} = 20$ s. The available coding schemes are the ones presented in Tab. 1.2 with the same corresponding probabilities p_k of memoryless channel model given in Tab. 1.3. The influence of the main input parameters is studied on the following performance parameters: the average resource utilization $\bar{U}$ and the average throughput per user $\bar{X}$.

1.8.1 Influence of the number of resources

First, we consider the influence of N_S, the number of available slots in the downlink subframe. Fig. 1.8 (a) shows that, for a given number of users and a given traffic load, the average resource utilization $\bar{U}$ decreases as N_S increases. As a direct consequence, Fig. 1.8 (b) shows that the average throughput per user linearly increases with N_S. When there are more resources, more slots are given to each active user (depending on the scheduling policy), which results in a better throughput. As a consequence, average transfer duration decreases and so does the average utilization of the cell. Obviously, increasing N_S is beneficial as it enables an increase in the average throughput per user for all scheduling policies. Note however that in a real-world dimensioning exercise, increasing N_S has a cost, since bandwidth requirements increases for rising the number of slots. This cost must then be balanced against the corresponding benefits.

1.8.2 Influence of the traffic load

We now study the effect of changing the traffic load (by varying the mean size of downloaded elements). We vary the mean size $\bar{x}_{on}$ from 2 to 10 Mbits, corresponding to $\rho \in [500; 2500]$ (see relation (1.4)). Performance curves on Fig 1.9 (a, b) show a similar behavior than the one observed in Section 1.7 when the number N of users in the cell varies. This common behavior reveals that traffic load is characterized by the combination of N, the population of the cell, and the traffic parameters expressed by $\rho = \bar{x}_{on}/\bar{t}_{off}$. Both have similar impact on performance. It is important to notice that, as the performance parameters depend on the ratio $\rho = \bar{x}_{on}/\bar{t}_{off}$, any couple of traffic characteristic giving the same value of ρ provides the same performance results.

The large range of values used for this study allows us to observe two regimes on the performance curves corresponding to: non-saturated and saturated systems. In the former, the average utilization increases linearly with the traffic load (with a slope that depends on the scheduling policy), and throughput decreases accordingly. As a matter of fact, transfer durations increase linearly with the size of the downloaded elements, and thus with the

traffic load. In the latter, resources are fully used and the system is under saturation. The utilization is close to 1 and the average throughput tends to attain very low values corresponding to infinite resource sharing. As shown in the figures, the curve slope of the non-saturated regime and thus the starting point of the saturated regime, strongly depend on the scheduling policy. These curves show that opportunistic scheduling postpones the transition between the two regimes to higher traffic loads.

1.9 Network design

We now provide and explain the use of some examples of graphs that can be instantaneously obtained with our analytical solution, and that can't even be thought of with simulators because of their prohibitive computation time. All these graphs correspond to the throughput fairness scheduling but can be drawn as easily for the two other scheduling policies (as well as for any alternative policy provided the average bit rates $\bar{m}(n)$ can be evaluated).

1.9.1 Performance graphs

We first draw 3-dimensional surfaces where performance parameters are function of, e.g., N, the number of users in the cell and ρ, the combination of traffic parameters (see relation (1.4)). For each performance parameter, the surface is cut out into level lines and the resulting 2-dimensional projections are drawn. The step between level lines can be arbitrarily chosen as a function of the required precision.

The average radio resource utilization of the WiMAX cell $\bar{U}$, and the average throughput per user $\bar{X}$ for any mobile in the system are presented in Fig. 1.10(a) and 1.10(b) (corresponding to the radio link characteristics presented in Section 1.7). Each graph is the result of several thousands of evaluation points (corresponding to varying input parameters). Obviously, any simulation tool or even any multi-dimensional Markov chain requiring numerical resolution, would have precluded the drawing of such graphs. These graphs allow to directly derive the corresponding performance parameter knowing the traffic load profile, i.e., the

couple (N, ρ). Measures on real systems can provide such parameters as they depend on the corresponding traffic profiles of the used applications (FTP, web, mail, etc.). These performance graphs can be used as easily and efficiently as the classical Erlang graphs employed for the dimensioning of telephone networks.

1.9.2 Dimensioning study

In this section, we show how our model can be advantageously applied for dimensioning issues. Once again, our modeling framework allows very fast computations, which in turn allows complex iterative dimensioning analyses. Each point of the following graphs now corresponds to multiple iterations of our model resolution, in order to find the optimal value of an input parameter (e.g., the number of users in the system) that respects a given QoS criterion.

As an example, Fig. 1.10(a) gives minimum values N_{min} of mobiles in the cell, in order to guarantee that the average radio utilization is over 50%. This kind of criterion allows operators to maximize the utilization of network resource in comparison with the traffic load of their customers. For a given traffic load profile and a given set of system parameters, the point of coordinates (N_S, ρ) in the graph is located between two level lines, and the level line with the higher value gives the optimal value of N_{min}.

Fig. 1.10(b) gives another example of compact and efficient dimensioning graphs, corresponding to the throughput offered to users. The QoS criterion requires that each mobile obtains a minimum average throughput for its transfers. Operators dimensioning issues usually require to find out an optimal value of N that satisfies a Maximum Instantaneous Rate (MIR) for users during their transfer, where the MIR simply corresponds to our average throughput per user $\bar{X}$. As explained before, the average throughput per user is a decreasing function of N. The higher the number of users, the higher the traffic load and the lower the throughput per user. We then have to find the maximum value N_{max} of users in the cell to guarantee the minimum throughput threshold. In Fig. 1.10(b), a given point (N_S, ρ) is located between two level lines. The line with the lower value gives N_{max}.

Note finally that these several dimensioning graphs can be used together for guaranteeing multiple QoS criteria. As an example, if we have a WiMAX cell configured to have $N_S = 450$ slots and a traffic profile given by $\rho = 300$ (e.g., $x_{on} = 1.2$ Mbits and $t_{off} = 20$ s), Fig. 1.10(a) gives $N_{min} = 55$, and Fig. 1.10(b) gives $N_{max} = 200$. As a consequence, the combination of these two graphs recommend to have a number of users $N \in [55; 200]$ to guarantee a reasonable utilization of the cell and to offer a minimum throughput to users.

1.10 Conclusion

As deployment of WiMAX networks is underway, need arises for operators and manufacturers to develop dimensioning tools. In this chapter, we have presented novel analytical models for WiMAX networks and elastic ON/OFF traffic. The models are able to derive Erlang-like performance parameters such as throughput per user or channel utilization. Based on a one-dimensional Markov chain and the derivation of average bit rates, whose expressions are given for three main scheduling policies (slot fairness, throughput fairness and opportunistic scheduling), our model is remarkably simple. The resolution of model provides closed-form expressions for all the required performance parameters at a click-speed. Therefore it will enable efficient and advanced dimensioning studies. The generic nature of model makes it flexible to be customized to scenario specific requirements. For example, the Markov chain can be adapted to any other scheduling policy since a general expression for the average bit rates is also given. Extensive simulations have validated the model's assumptions. The accuracy of the model is illustrated by the fact that, for all simulation results, maximum relative errors do not exceed 10%. Even if the traffic and channel assumptions are relaxed, analytical results still match very well with simulations that show the robust nature of our model.

References

[1] IEEE 802.16e: IEEE 802.16e Task Group (Mobile WirelessMAN) http://www.ieee802.org/16/tge/

[2] IEEE Std. 802.16: IEEE Standard for local and metropolitan area networks - Part 16: Air Interface for Fixed Broadband Wireless Access Systems, 2004.

[3] IEEE Std. 802.16e: Draft IEEE std 802.16e/D9. IEEE Standard for local and metropolitan area networks - Part 16: Air Interface for Fixed Broadband Wireless Access Systems. Amendment 2: Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands, 2005.

[4] B. Baynat, K. Boussetta, P. Eisenmann and N. Ben Rached, "Towards an Erlang-Like formula for the performance evaluation of GPRS/EDGE networks with finite-length sessions", 3rd IFIP-TC6 Networking Conference, May 2004.

[5] B. Baynat and P. Eisenmann, "Towards an Erlang-Like formula for GPRS/EDGE network engineering", IEEE International Conference on Communications (ICC), June 2004.

[6] B. Baynat, K. Boussetta, P. Eisenmann and N. Ben Rached, "A discrete-time Markovian model for GPRS/EDGE radio engineering with finite-length sessions traffic", International Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS04), July 2004.

[7] A. Berger and Y. Kogan, "Dimensioning bandwidth for elastic traffic in high-speed data networks", IEEE/ACM Transactions on Networking, vol.8, pages 643–654, October 2000.

[8] C. Cicconetti, L. Lenzini, E. Mingozzi and C. Eklund, "Quality of Service Support in IEEE 802.16 Networks", IEEE Network, March 2006.

[9] T. O. Engset, "On the calculation of switches in an automatic telephone system", Tore Olaus Engset: The man behind the formula - First published in Norwegian (1915), 1998.

[10] A. Feldmann, A. C. Gilbert, P. Huang and W. Willinger, "Dynamics of IP traffic: A study of the role of variability and the impact of control", ACM SIGCOMM Computer Communication Review, October 1999.

[11] S. Ben Fredj, T. Bonald, A. Proutiere, G. Regnie and J. Roberts, "Statistical bandwidth sharing: A Study of congestion at flow level", ACM Special Interest Group on Data

Communications (Sigcomm), August 2001.

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

[13] S. Borst, "User-level performance of channel-aware scheduling algorithms in wireless data networks", IEEE Infocom, April 2003.

[14] D. Heyman, T. Lakshman and A. Neidhardt, "New method for analyzing feedback protocols with applications to engineering web traffic over the internet, ACM Sigmetrics, June 1997.

[15] C. Y. Huang, H.-H. Juan, M.-S. Lin and C.-J. Chang, "Radio Resource Management of Heterogeneous Services in Mobile WiMAX Systems", IEEE WCNC, February 2007.

[16] G. Kulkarni, S. Adlakha, M. Srivastava, "Subcarrier Allocation and Bit Loading Algorithms for OFDMA-Based Wireless Networks", IEEE Trans. on Mobile Computing, December 2005.

[17] T. Kwon, H. Lee, S. Choi, J. Kim, D.-H. Cho, S. Cho, S. Yun, W.-H. Park and K. Kim, "Design and Implementation of a Simulator Based on a Cross-Layer Protocol between MAC and PHY Layers in a WiBro Compatible IEEE 802.16e OFDMA System", IEEE Communication Magazine, December 2005.

[18] H. Lee, T. Kwon, D. H. Cho, G. Lim and Y. Chang, "Performance Analysis of Scheduling Algorithms for VoIP Services in IEEE 802.16e Systems", Vehicular Technology Conference (VTC), 2006.

[19] S. Liu and J. Virtamo, "Performance Analysis of Wireless Data Systems with a Finite Population of Mobile Users", 19th International Teletraffic Congress, 2005.

[20] M. Maqbool, M. Coupechoux, P. Godlewski, "Comparative Study of Reuse Patterns for WiMAX Cellular Networks", Technical Report Telecom Paris, 2007.

[21] D. Niyato and E. Hossain, "A queuing-theoretic and optimization-based model for radio resource management in IEEE 802.16 broadband networks", IEEE Transactions on Computers, vol.55 pages 1473–1488, 2006.

[22] G. Nogueira, "Methodes analytiques pour le dimensionnement des reseaux cellulaires, PhD thesis at UMPC - Paris 6 (2007), http://www-rp.lip6.fr/∼nogueira/pdf/theseGN.pdf.zip

[23] K. Ramadas and R. Jain, "WiMAX System Evaluation Methodology (Wimax Forum),

January 2007.

[24] M. Settembre, M. Puleri, S. Garritano, P. Testa, R. Albanese, M. Mancini and V. Lo Curto, "Performance Analysis of an Efficient Packet-Based IEEE 802.16 MAC Supporting Adaptive Modulation and Coding, 7th IEEE ISCN, 2006.

[25] V. Singh and V. Sharma, "Efficient and Fair Scheduling of Uplink and Downlink in IEEE 802.16 OFDMA Networks, IEEE WCNC, April 2006.

[26] J. Sun, Yanling and H. Zhu, "Quality of Service Scheduling for 802.16 Broadband Wireless Access Systems", IEEE VTC Spring, May 2006.

[27] A. Vinel, Y. Zhang, M. Lott and A. Tiurlikov, "Performance analysis of the random access in IEEE 802.16", IEEE PIMRC, September 2005.

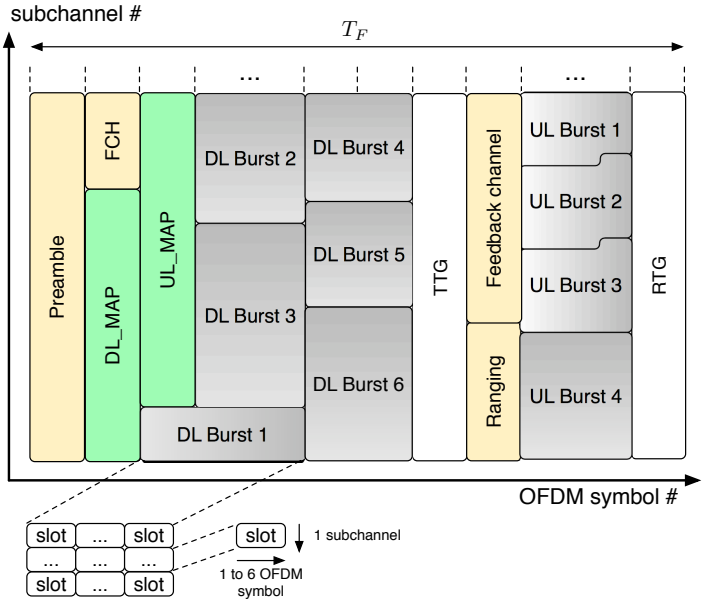


Figure 1.1: TDD frame structure.

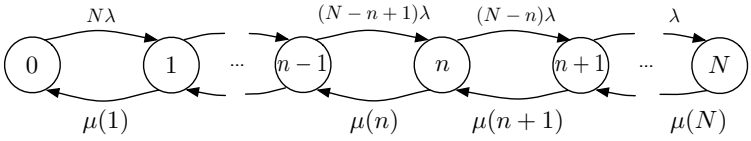


Figure 1.2: General CTMC with variable departure rates.

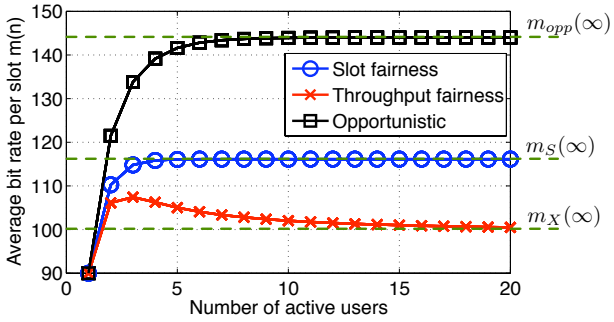
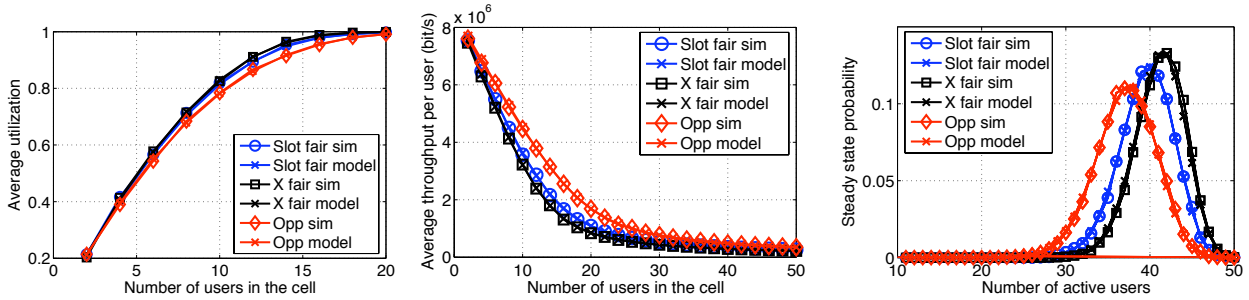


Figure 1.3: $\bar{m}(n)$ asymptotic behavior.



(a) Average utilization. (b) Average throughput per user. (c) Stationary probabilities for $N = 50$.

Figure 1.4: Performance validation for the three scheduling policies with $\bar{x}_{on} = 3$ Mbits and $\bar{t}_{off} = 3$ s.

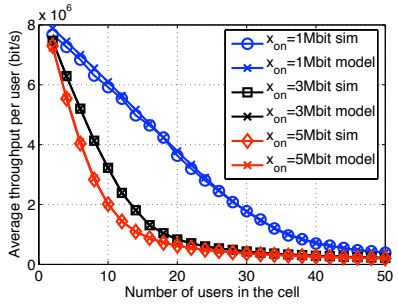


Figure 1.5: Average throughput per user for different loads.

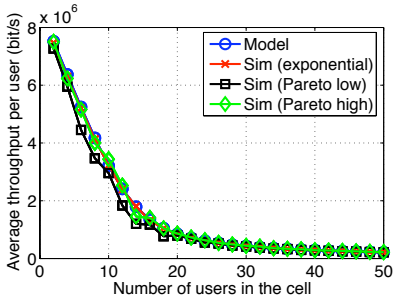


Figure 1.6: Average throughput per user for different traffic distributions.

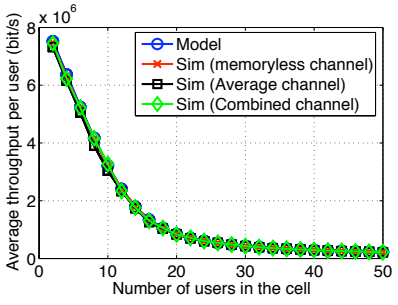
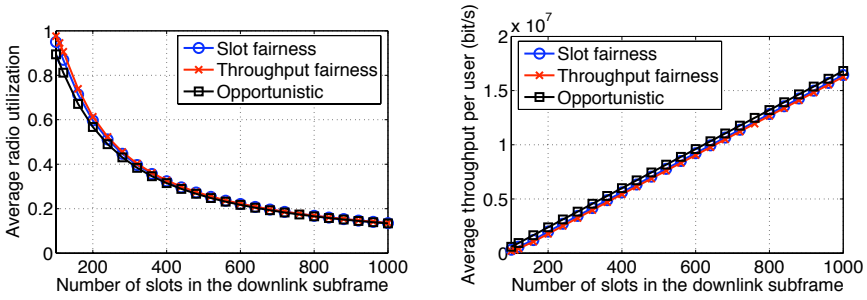
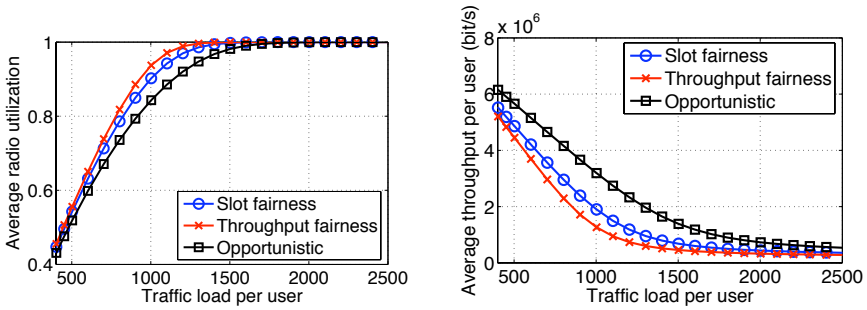


Figure 1.7: Average throughput per user for different channel models.



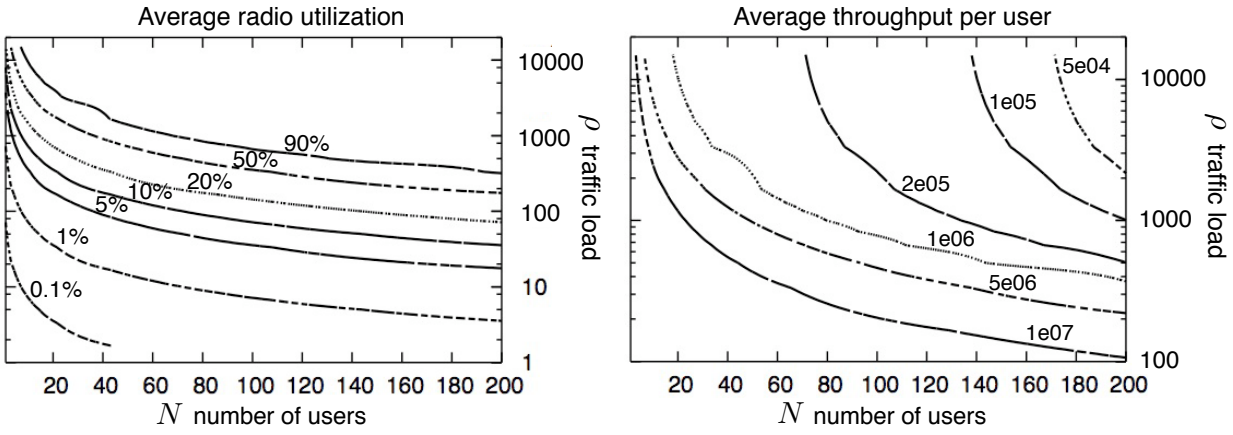
(a) Average utilization $\bar{U}$. (b) Average throughput per user $\bar{X}$.

Figure 1.8: Influence of the number of resources.



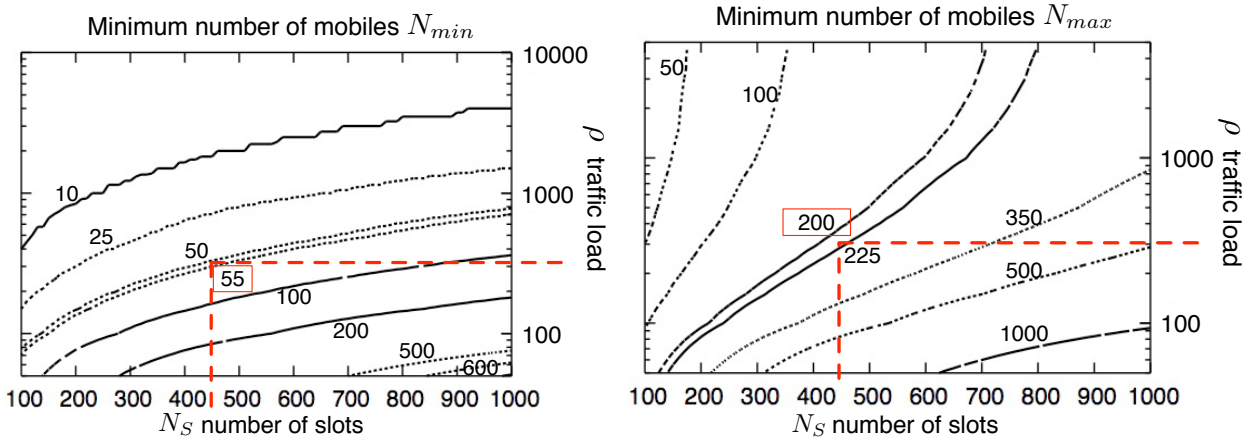
(a) Average utilization $\bar{U}$. (b) Average throughput per user $\bar{X}$.

Figure 1.9: Influence of the traffic load.



(a) Average utilization $\bar{U}$. (b) Average throughput per user $\bar{X}$.

Figure 1.10: Performance graphs



(a) Dimensioning the minimum value of N for having $\bar{U} \geq 50\%$.

(b) Dimensioning the maximum value of N for having $\bar{X} \geq 50$ Kbps per user.

Figure 1.11: Dimensioning graphs