An Efficient Analytical Model for the Dimensioning of WiMAX Networks Supporting Multi-profile Best Effort Traffic

Bruno Baynat, Sébastien Doirieux, Masood Maqbool, Marceau Coupechoux

An Efficient Analytical Model for the Dimensioning of WiMAX Networks Supporting Multi-profile Best Effort Traffic

Bruno Baynat, Sébastien Doirieux

LIP6 - UPMC Paris Universitas - CNRS, Paris, France

Masood Maqbool, Marceau Coupechoux

Télécom ParisTech and CNRS LTCI, Paris, France

Abstract

This paper tackles the challenging task of developing a simple and accurate analytical model for performance evaluation of WiMAX networks. The need for accurate and fast-computing tools is of primary importance to face complex and exhaustive dimensioning issues for this promising access technology. In this paper, we present a generic Markovian model developed for three usual scheduling policies (slot sharing fairness, throughput fairness and opportunistic scheduling) that provides closed-form expressions for all the required performance parameters instantaneously. We also present and evaluate the performance of a fourth policy, called throttling policy, that limits the maximum user throughput and makes use of the Maximum Sustained Traffic Rate (MSTR) parameter foreseen by the standard. At last, we extend these studies to multi-profile traffic patterns. The proposed models are compared in depth with realistic simulations that show their accuracy and robustness regarding the different modeling assumptions. Finally, the speed of our analytical tools allows us to carry on dimensioning studies that require several thousands of evaluations, which would not be tractable with any simulation tool.

Key words: WiMAX, performance evaluation, dimensioning, analytical models, multi-profile traffic, best effort

Contents

1. Introduction

Candidate for 4G, WiMAX (Worldwide Interoperability for Microwave Access) is a broadband wireless access technology which is based on IEEE standard 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, 802.16m, is currently under definition to provide even higher efficiency. In addition, 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).

A number of services such as voice, video and web are to be offered by WiMAX networks. Considering the web services, the users may generate traffic of different profiles (characterized by the volume of data generated and reading time). They may also have to respect a QoS parameter associated with best effort service in the standard: the *Maximum Sustained Traffic Rate* (MSTR). As defined in [3] (section 11.13.6), this parameter is not a guaranteed rate but an upper bound on the throughput achieved by a mobile. 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 able to account for these possibilities. Literature on WiMAX performance evaluation is constituted of two sets of papers: i) packet-level simulations that

precisely implement system details and scheduling schemes; ii) analytical models and optimization algorithms that derive performance metrics at user-level.

In the former set, [19] and [12] are interesting because they investigate different QoS support mechanisms proposed in the standard. In addition, studies of the performance of multi-profile internet traffic have been proposed in both [25] and [23]. Authors of [25] evaluated the throughput performance in a WiMAX cell while considering the number of users, the modulation schemes they may use and the data rate they require, using system level simulations. They also introduced a notion similar to MSTR: the *Mean Information Rate* (MIR) and observed the impact of different MIR values on the traffic performance. In [23], a measurement based procedure has been adopted to evaluate the performance of fixed WiMAX network in presence of multi-profile best effort traffic.

Among the latter set of papers, [26] provided an analytical model for studying the random access scheme of IEEE 802.16d. Niyato and Hossain [21] formulated the bandwidth allocation of multiple services with different QoS requirements by using linear programming. They also proposed performance analysis, first at connection level, 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 rely on multi-dimensional Markovian model that requires numerical resolutions. Finally, authors of [27] presented the mathematical expressions to calculate the blocking probabilities of a mixed WiMAX-WiFi system. They considered users who generate voice/data traffic and focused on the admission control aspect of the network.

Not specific to WiMAX systems, generic analytical models for performance evaluation of cellular networks with varying channel conditions have been proposed in [11, 10, 20]. 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 propagation models, mobility models or spatial distribution of users in a cell. These papers implicitly consider 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.

In this paper, 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 [11, 10, 20], our model is adapted to WiMAX systems' assumptions and is generic enough to integrate any appropriate scheduling policy. Moreover, our approach makes it possible to consider the so-called "outage" situation. A user experiences an outage, if at a given time radio conditions are so bad that it cannot transfer any data and is thus not scheduled.

We first consider three *full-capacity* policies which aim at sharing the whole resource, i.e., all slots of each frame, among the active users: *slot sharing fairness*, *instantaneous throughput fairness*, and *opportunistic*. Then, we consider a *throttling* scheduling policy which limits the attained throughput of each user to a given value. This policy allows us to take into account the aforementioned MSTR in our model. For each policy, we develop closed-form expressions for all performance metrics. We also provide extensions of our model to take into account multi-profile web traffic in mobile WiMAX networks.

The rest of paper is organized as follows. System description including specific WiMAX network details concerning our analytical model is provided in Section 2. Section 3 presents the generic analytical model and the assumptions it stands on. The model is adapted to the three full-capacity scheduling policies in Section 4 and to the throttling policy in Section 5. Section 6 details the multi-profile traffic extensions for both kinds of scheduling policy. Validation and robustness of model are discussed in Section 7. Lastly, Section 8 gives an example of WiMAX dimensioning process using our model.

2. 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 resources (slots), the access technique, and the different Modulation and Coding Schemes (MCS) are presented. Finally we also introduce the different scheduling policies considered in this work.

2.1. WiMAX Standard

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 [18]. 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 permutations, they take advantage of channel diversity among subchannels. Adaptive modulation and coding (AMC) is a type of adjacent permutation, it 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 [3]. An example of a WiMAX TDD frame is shown in Fig. 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, although our model is valid for any permutation scheme. Note however that a slot always carries 48 subcarriers whatever the type of used subcarrier permutation. In the DL sub-frame, a first part contains a Preamble, a Frame Control Header (FCH), a UL_MAP and a DL_MAP. The preamble is used for synchronization. The FCH provides length and encoding of two MAP messages and information about usable subchannels. Finally, in the MAP messages reside the data mapping for users. Their sizes depend on the number of scheduled users in the frame.

One of the important features of IEEE 802.16e is link adaptation: using different MCS enables 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 for each mobile wishing to transmit (i.e., active mobile) according to its signal to interference plus noise ratio (SINR). However, note that when the SINR is too low, no data can be transmitted without error. This situation is called *outage*.

At last, let us highlight that WiMAX networks are to carry all sorts of applications. To answer the various QoS needs of these applications, several service classes have been defined in the WiMAX standard ranging from high QoS-guaranteed classes supporting real-time applications to a best effort class mostly for WEB services. In this study, we only consider traffic from the best effort service class. How to integrate the other service classes into the model we provide here will be addressed in future works.

2.2. Scheduling Policies

The scheduling algorithm is responsible for allocating the radio resources of every frame to active users. In wireless networks, scheduling may take into account their radio link quality. No scheduling policy is recommend by the WiMAX standard so, in this work, we consider four generic schemes.

First we consider three *full-capacity* policies which aim at sharing the whole resource, i.e., all slots of each frame, among the active users: 1) the *slot sharing fairness* scheduling equally divides slots between active users, regardless of their radio conditions, 2) the *throughput fairness* scheduling ensures that all active users achieve the same instantaneous throughput, and 3) the *opportunistic* scheduling gives all the resources to the active users with the best channel conditions (i.e., the best MCS).

Then we consider a *throttling* scheduling policy which aims at limiting the attained throughput of each active user to a value called *MSTR* (*Maximum Sustained Traffic Rate*). As opposed to the former set of policies, if there are still resources left in the frame after ensuring that each active user attains his maximum throughput, these resources go unused.

2.3. Notations

Let us now define the notations concerning the WiMAX system used in this article:

- 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 duration of one TDD frame: $T_F = 5$ ms.
- Radio channel states are denoted MCS_k, $0 \leq k \leq K$, 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 a mobile using MCS_k. Let us recall that the bit rate per slot is independent of the permutation method and is thus constant for a given MCS. For the particular case of outage, $m_0 = 0$.

3. WiMAX Generic Analytical Model

In this section, we develop a generic analytical model able to account for any scheduling policy. Then, in the two following sections, we will see how to adapt this generic model either to any of the full-capacity policies (Section 4) or to the throttling policy (Section 5).

3.1. Modeling Assumptions

We consider a single WiMAX cell handling data traffic belonging only to the best effort service class of WiMAX. This study targets the analysis of the bottleneck, i.e., the radio link, and focuses on the downlink part which is assumed to be a critical resource in asymmetric data traffic. However, nothing prevents using the model to characterize the uplink part.

The development of our analytical model stands on several assumptions related either to the system, the channel or the traffic. All of them will be discussed in Section 3.3, and, as will be shown in that section, most of them can be relaxed, if necessary, by slightly modifying the basic model.

3.1.1. 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 equal N_S.
2. We assume that the number of mobiles that can simultaneously be in active transfer is not limited. As a consequence, any connection demand will be accepted and no blocking can occur.
3. We consider that each mobile has unlimited transmission capacities. So, at any given time, if there is only one active user, he can use all the available slots of the frame for his transfer if allowed by the scheduler.

3.1.2. Channel assumption

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 a feedback channel estimation on a frame by frame basis, and thus, the base station can change its coding scheme at 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).

3.1.3. Traffic assumptions

5. All the users have the same traffic characteristics. This assumption is relaxed in Section 6 where we propose multi-profile traffic extensions of the model.
6. We do not take handover into account.
7. We assume that there is a fixed number N of mobiles sharing the available bandwidth of the cell.

8. As we only consider the best effort service class of WiMAX, 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). 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 are characterized by their duration.
9. We assume that both ON sizes and OFF durations are exponentially distributed. We denote by $\overline{x_{on}}$ the average size of ON data volumes (in bits) and $\overline{t_{off}}$ the average duration of OFF periods (in seconds).

In short, we assume that N users are generating infinite-length ON/OFF best effort traffic with the same traffic profile $(\overline{x_{on}}, \overline{t_{off}})$.

3.2. Generic Analytical Model

3.2.1. Model description

A first attempt for modeling this system would be to develop a multidimensional Continuous Time Markov Chain (CTMC). A state $(n_0, ..., n_K)$ of this chain would be a precise description of each current number n_k of mobiles using coding scheme MCS_k, $0 \leq k \leq K$ (i.e., 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 this complexity problem, we aggregate the state description of the system into a single dimension n, representing the total number of concurrent active mobiles, regardless of the coding scheme they use. The resulting CTMC is thus made of $N + 1$ states as shown in Fig 2.

- A transition out of a generic state n to state $n + 1$ occurs when a mobile in OFF period initiate a data transfer. This "arrival" transition corresponds to one mobile 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:

$$\lambda = \frac{1}{\overline{t_{off}}}, \tag{1}$$

- 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 performed with a generic rate $\mu(n)$ corresponding to the total departure rate when n mobiles are active.

Obviously, the main difficulty of the model resides in estimating the aggregate departure rates $\mu(n)$ that strongly depend on the chosen scheduling policy. This will be done in Sections 4 and 5, for the four generic policies we consider in this paper.

3.2.2. Performance parameters

Provided that the departure rates $\mu(n)$ can be conveniently estimated, the steady-state probabilities $\pi(n)$ can easily be derived from the birth-and-death structure of the Markov chain as:

$$\pi(n) = \Big(\prod_{i=1}^{n} \frac{(N - i + 1)\lambda}{\mu(i)} \Big) \pi(0), \tag{2}$$

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

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

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

$$\overline{Q} = \sum_{n=1}^{N} n\, \pi(n). \tag{3}$$

$\bar{D}$, the mean number of departures (i.e., mobiles completing their transfer) by unit of time, is obtained as:

$$\bar{D} = \sum_{n=1}^{N} \mu(n)\,\pi(n). \tag{4}$$

From Little's law, we thus derive the average duration $\bar{t_{on}}$ of an ON period (duration of an active transfer):

$$\bar{t_{on}} = \frac{\bar{Q}}{\bar{D}}, \tag{5}$$

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

$$\bar{X} = \frac{\bar{x_{on}}}{\bar{t_{on}}}. \tag{6}$$

Lastly, we can derive the average utilization $\bar{U}$ of the TDD frame. However the expression of this parameter strongly depends on the considered policy (see Sections 4 and 5).

3.3. Discussion of the Modeling Assumptions

Our Markovian model is based on the system, radio channel and traffic assumptions presented in Section 3.1. We now discuss these assumptions one by one (item numbers are related to the corresponding assumptions), evaluate their accuracy, and provide, if possible, extensions and generalization propositions.

1. As described in Section 2, 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 sizes of these MAPs, and as a consequence the number N_S of available slots for downlink data transmissions, depend 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 easily be integrated into the model by replacing N_S by $N_S(n)$ (and N_S^n by $\prod_{i=1}^{n} N_S(i)$) in the expressions of the departure rates $\mu(n)$, of the steady states probabilities $\pi(n)$ and of the average utilisation ratio $\bar{U}$ (relations (7, 28) for the full-capacity policies or relations (35, 36, 38) for the throttling policy).
2. In order to consider the possibility of an admission control, a limit n_{max} on the total number of mobiles allowed to be in active transfer simultaneously, can easily be introduced in the model. The corresponding Markov chain shown in Fig. 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 [6].
3. In some cellular networks (e.g., GPRS/EDGE), the mobiles have limited transmission capabilities because of hardware considerations. This constraint defines a maximum throughput the network interface can reach or a maximum number of resource units that can be used by the mobiles. Such limitations add a slight complexity in the model development as one single mobile may only use a limited number of slots. This characteristic has been introduced in the case of (E)GPRS networks [6, 22] and can be applied to WiMAX networks by simply modifying the departure rates of the first states of the Markov chain (e.g., if d is the maximum number of slots a mobile can use, replace N_S by $\min(nd, N_S)$ in the same relations as those listed in point 1).
4. The 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 the stationary probabilities of using the different coding schemes and thus does not explicitly take into account the radio channel dynamics. This approach is authenticated through simulations in Section 7.
5. All mobiles in the considered system have statistically the same traffic characteristics. As stated before, this assumption is relaxed in Section 6 were multi-profile traffic extensions are provided for both kinds of scheduling policies.
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 indirectly taken into account in the channel model by means of radio conditions variation.

7. Poisson processes are commonly 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 model. In addition, the finite population assumption is used typically for network planning when geo-marketing data allows predicting 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 that 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 an infinite length ON/OFF session traffic. In the context of (E)GPRS networks, we have studied similar processor sharing systems [7, 5], and provided extensions to finite length sessions, where each mobile generates an ON/OFF traffic during a session and does not generate any traffic during an inter-session. We have shown in [7] that a very simple transformation of the traffic characteristics that increases the OFF periods by a portion of the inter-session period, enables us to transform the resulting bi-dimensional model into a linear Erlang-like model, with very good accuracy. The accuracy of this transformation is related to the insensibility of the average performance parameters with regards to the traffic distributions, that comes from the processor sharing policy (see the next point). An equivalent transformation can be applied to the WiMAX model developed here. However, because of the specificity of the scheduling policy in WiMAX networks, the system is no longer processor sharing. Even if, in some cases, it can be considered as a generalized processor sharing system, we were not able to prove that the transformation (from finite-length to infinite length sessions) is exact, but several experiments have shown us that it is at least a very good approximation.
9. Memoryless traffic distributions are strong assumptions that have been validated by several theoretical results. Several works on insensitivity (see, e.g., [9, 11, 17]) have shown (for systems fairly similar to the one we are studying) that the average performance parameters are insensitive to the distribution of ON and OFF periods. In Section 7, we present 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 insensitivity still holds or is at least a good approximation. Thus, memoryless distributions are the most convenient choices to model the traffic.

4. Full-capacity Policy Modeling

In this section we adapt the generic model presented in Section 3, by providing the expressions of the aggregated departure rates $\mu(n)$ in the case of a full-capacity policy. The resulting model has been presented in [8].

4.1. Full-capacity Policy

A full-capacity policy, as specified in Section 2.2, is a scheduling policy that aims at always sharing the whole resource between the users. So, as long as there is at least one active mobile (i.e., a mobile currently in active transfer) that is not in outage, all the slots of the current frame are given to this user and no resources go unused.

In this study, we consider three full-capacity policies coresponding to three specific scheduling schemes:

- The slot sharing fairness policy equally shares all slots of each frame between the active users that are not in outage. Obviously, since users with better MCS make better uses of these slots, they achieve greater instantaneous throughputs.
- The instantaneous throughput fairness policy divides the resource in order to provide the same instantaneous throughput to all active users not in outage. This policy allows mobiles using at a given time-step a MCS with a low bit rate, to obtain proportionally more slots than mobiles using a MCS with a high bit rate.
- The opportunistic policy gives all the resources to active users having the highest transmission bit rate, i.e., the best MCS. This policy ensures the most efficient use of each frame at the cost of unfairness between the users.

It is of importance to note that the two last policies correspond to two opposite borderline cases. Indeed, while the instantaneous throughput fairness policy totally favors fairness between the active mobiles over an efficient use of the ressource, the opportunistic policy does the exact opposite. Finally, the slot sharing policy can be seen as a trade-off between these two policies.

4.2. Departure Rates

To accurately estimate the average departure rates $\mu(n)$ of this model, we first express them as follows:

$$\mu(n) = \frac{\bar{m}(n)\, N_S}{\bar{x}_{on}\, T_F}, \tag{7}$$

where $\bar{m}(n)$ is the average number of transmitted bits 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. $\bar{m}(n)$ is also strongly dependent on n, because the average number of bits per slot must be estimated by considering all possible distributions of the n mobiles between the $K+1$ possible coding schemes (including outage). Finally, $\bar{m}(n)$ also depends on the scheduling policy, as the policy decides on the numbers of slots given to each mobile with regard to the coding scheme they use.

4.2.1. Generic Average Bit Rates

In order to illustrate the derivation of the generic average numbers of bits per slot $\bar{m}(n)$, we first consider a situation with 2 active mobiles (denoted M1 and M2) in a system with 2 MCS ($K = 2$) and no outage, 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 by $\bar{m}(n_1, n_2)$ the average number of bits per slot in the TDD frame for one particular configuration having n_1 mobiles using MCS_1 and n_2 mobiles using MCS_2 ($n_1 + n_2 = 2$). There are 3 possible configurations:

- the 2 mobiles uses MCS_1 and thus none uses MCS_2. This configuration occurs with a probability $p_1 p_1$. Whatever the scheduling policy, the corresponding average number of bits per slot $\bar{m}(2,0)$ is obviously:

$$\bar{m}(2,0) = m_1; \tag{8}$$

- the 2 mobiles use MCS_2 and thus none uses MCS_1. Similarly, with a probability $p_2 p_2$, we have:

$$\bar{m}(0,2) = m_2; \tag{9}$$

- 1 mobile uses MCS_1 and the other uses MCS_2. This configuration can correspond to 2 distinct distributions of the 2 mobiles: M1 uses MCS_1 and M2 uses MCS_2, or M1 uses MCS_2 and M2 uses MCS_1. The associated probability is $2p_1p_2$, as both distributions have equal probabilities. The corresponding average number of bits per slot $\bar{m}(1,1)$ can thus be expressed as:

$$\bar{m}(1,1) = m_1 x_1(1,1) + m_2 x_2(1,1), \tag{10}$$

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

We finally express the average number of bits per slot when there are 2 active mobiles in the system as:

$$\bar{m}(2) = \sum_{n_1=0}^{2} \bar{m}(n_1, 2-n_1) \binom{2}{n_1} p_1^{n_1} p_2^{2-n_1}, \tag{11}$$

where $\binom{2}{n_1}$ is a binomial coefficient that gives the number of distributions corresponding to a same configuration (with n_1 mobiles using MCS_1 and $2 - n_1$ mobiles using MCS_2).

As a generalization, one can convince oneself easily that the average number of bits per slot, $\bar{m}(n)$, when there are n active users, is expressed as follows:

$$\bar{m}(n) = \sum_{\substack{(n_0,...,n_K)=(0,...,0)| \\ n_0+...+n_K=n \\ n_0 \neq n}}^{(n,...,n)} \bar{m}(n_0,...,n_K) \binom{n}{n_0,...,n_K} \left(\prod_{k=0}^{K} p_k^{n_k} \right), \tag{12}$$

with

$$\bar{m}(n_0,...,n_K) = \sum_{k=1}^{K} m_k \, n_k \, x_k(n_0,...,n_K), \tag{13}$$

where $\binom{n}{n_0,...,n_K}$ is the multinomial coefficient and $x_k(n_0,...,n_K)$ is the proportion of resource given to one mobile using MCS_k, when the current distribution of the n mobiles among the $K+1$ coding schemes (including outage) is $(n_0,...,n_K)$. Let us emphasize that this expression has a $O(n^K)$ complexity, where K, the number of different coding schemes, is usually low. Section 4.2.3 will show that this complexity can be drastically reduced without any significant impact on the accuracy of the $\bar{m}(n)$ values.

4.2.2. Specific Policies

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

Slot sharing fairness

With the slot fairness policy, at each time-step the scheduler equally shares the N_S slots among the active users that are not in outage. As a result, if we consider a particular distribution $(n_0,...,n_K)$ of the n active users ($n = \sum_{k=0}^{K} n_k$), each of the $n - n_0$ users not in outage receives an equal portion of the whole resource. The proportion $x_k(n_0,...,n_K)$ of the resource granted to a mobile using MCS_k, is thus given by:

$$x_k(n_0,...,n_K) = \begin{cases} \frac{1}{n-n_0} & \text{if } k \neq 0 \text{ and } n \neq n_0 \\ 0 & \text{else} \end{cases} \tag{14}$$

By replacing these proportions in generic expression (12), the average number of bits per slot, $\bar{m}(n)$, when there are n active users, becomes:

$$\bar{m}(n) = \sum_{\substack{(n_0,...,n_K)=(0,...,0)| \\ n_0+...+n_K=n \\ n_0 \neq n}}^{(n,...,n)} \frac{n!}{n-n_0} \left(\sum_{k=1}^{K} m_k n_k \right) \left(\prod_{k=0}^{K} \frac{p_k^{n_k}}{n_k!} \right). \tag{15}$$

It is of interest to note that the expression of the average numbers of bits per slot can be greatly simplified if we don't consider outage. Indeed, in that case, an active mobile can always receive data. So, the proportion of the resource that is associated to mobiles using MCS_k, is then constant for any k and for any possible distribution $(n_1,...,n_K)$ of the n mobiles among the K coding schemes, and equals $\frac{1}{n}$. These constant proportions when replaced in generic expression (12) lead, after a few simplifications, to the drastically simplified expression:

$$\bar{m}(n) = \sum_{k=1}^{K} m_k p_k = \bar{m}. \tag{16}$$

This nice and very simple expression shows us that, when there is no outage, the average numbers of bits per slot $\bar{m}(n)$ and, as a consequence, the average departure rates of the CTMC, are constant. As a result, in this special case (i.e., slot sharing fairness policy and no possible outage) our model becomes equivalent to the well-known Engset model [15].

Instantaneous throughput fairness

The objective of the instantaneous throughput fairness policy is to ensure that all active mobiles not in outage get the same instantaneous throughput. If an active mobile using coding scheme MCS_k obtains a proportion $x_k(n_0, ..., n_K)$ of the resource, its resulting instantaneous throughput will be proportional to $m_k x_k(n_0, ..., n_K)$. As a consequence, in order to respect instantaneous throughput fairness between active users not in outage, the $x_k(n_0, ..., n_K)$ must be such that:

$$m_k x_k(n_0, ..., n_K) = C \text{ for } k \neq 0, \tag{17}$$

where C is a constant such that $\sum_{k=1}^{K} n_k x_k(n_0, ..., n_K) = 1$, thus:

$$C = \frac{1}{\sum_{k=1}^{K} \frac{n_k}{m_k}}. \tag{18}$$

By replacing the proportions $x_k(n_0, ..., n_K)$ in generic expression (12), the average number of bits per slot, $\bar{m}(n)$, when there are n active users, becomes:

$$\bar{m}(n) = \sum_{\substack{(n_0, ..., n_K) = (0, ..., 0) | \\ n_0 + ... + n_K = n \\ n_0 \neq n}}^{(n,...,n)} \frac{(n - n_0)\, n! \prod_{k=0}^{K} \frac{p_k^{n_k}}{n_k!}}{\sum_{k=1}^{K} \frac{n_k}{m_k}}. \tag{19}$$

Unlike the previous scheduling scheme, here there is no significant simplification of the formula when we rule out the possibility of outage.

Opportunistic scheduling

Finally, we study the case of the opportunistic policy. Without loss of generality, we assume, in this section, that the coding schemes are classified in increasing order: $m_0 < m_1 < ... < m_K$. And even if it is still possible to derive the average numbers of bits per slot from generic expression (12), we prefer to give here a more intuitive, yet strictly equivalent, derivation.

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 enabling a higher transmission rate (i.e., MCS_j with $j > i$). As a matter of fact, $\alpha_i(n)$ corresponds to the probability that the scheduler grants at a given time-step all the resource to the mobiles that use MCS_i. As a consequence, we can express the average number of bits per slot when there are n active users as:

$$\bar{m}(n) = \sum_{i=1}^{K} \alpha_i(n) m_i. \tag{20}$$

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:

$$p_{\leq i}(n) = \left(1 - \sum_{j=i+1}^{K} p_j\right)^n. \tag{21}$$

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

$$p_{=i}(n) = 1 - \left(1 - \frac{p_i}{\sum_{j=0}^{i} p_j}\right)^n. \tag{22}$$

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

$$\alpha_i(n) = p_{=i}(n)\, p_{\leq i}(n). \tag{23}$$

Lastly, let us note that there is no change in the previous expressions whether we consider the possibility of outage or not. (In the latter, setting $p_0 = 0$ is sufficient to obtain the desired $\overline{m}(n)$.)

4.2.3. Analytical asymptotic study

In the derivation of the average numbers of bits per slot, we can observe the asymptotic behaviors of the $\overline{m}(n)$ functions. Fig. 3 shows the evolution of $\overline{m}(n)$ when n increases for the three studied scheduling policies. We can notice that the three resulting functions $\overline{m}(n)$ rapidly tend to different asymptotes as n, the number of active users, increases. We thus derive in the following subsections the analytical expressions of these asymptotes for each scheduling policy. Note that one can benefit from these quick asymptotical behaviors to avoid the calculation of the $\overline{m}(n)$ for large values of n (e.g., by replacing, after a threshold, the exact value by the corresponding asymptote value).

Slot fairness asymptote

As the number of active users grows, the proportion of mobiles using MCS_k tends to p_k. If we denote by n_k the number of such mobiles, when $n \to \infty$, we have $N_k \sim p_k\, n$. In the case of slot fairness scheduling, the resource is equally shared between the $n - N_0$ mobiles that are not in outage. The limiting value of the average number of bits per slot is thus given by:

$$m(\infty) = \lim_{n\to\infty} \overline{m}(n) = \lim_{n\to\infty} \sum_{k=1}^{K} m_k \frac{N_k}{n - N_0} = \frac{\sum_{k=1}^{K} m_k p_k}{1 - p_0}. \tag{24}$$

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 $N_k \sim p_k\, n$. Every such mobile obtains a proportion x_k of the resource such that $\sum_{k=1}^{K} N_k x_k = 1$. In order to respect the fairness of the scheduling policy, these proportions must satisfy the following relation:

$$m_k x_k = C = \frac{1}{\sum_{k=1}^{K} \frac{N_k}{m_k}} \text{ for any } k \neq 0. \tag{25}$$

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:

$$m(\infty) = \lim_{n\to\infty} \overline{m}(n) = \lim_{n\to\infty} \sum_{k=1}^{k} m_k N_k x_k = \frac{1 - p_0}{\sum_{k=1}^{K} \frac{p_k}{m_k}}. \tag{26}$$

Opportunistic scheduling asymptote

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

$$m(\infty) = \lim_{n\to\infty} \overline{m}(n) = m_K. \tag{27}$$

4.3. Performance parameters

From the expression of the full-capacity departure rates $\mu(n)$ (equation (7)), we now obtain the steady-states probabilities $\pi(n)$ as follows:

$$\pi(n) = \frac{N!}{(N-n)!} \frac{T_F^n \rho^n}{N_S^n \prod_{i=1}^{n} \bar{m}(i)} \pi(0), \tag{28}$$

where

$$\rho = \frac{\bar{x}_{on}}{\bar{t}_{off}}, \tag{29}$$

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

The performance parameters are derived from the steady-state probabilities as shown in Section 3.2.2. For example, the average throughput $\bar{X}$ obtained by each mobile in active transfer can be expressed as:

$$\bar{X} = \frac{N_S}{T_F} \frac{\sum_{n=1}^{N} \bar{m}(n)\,\pi(n)}{\sum_{n=1}^{N} n\,\pi(n)}. \tag{30}$$

The average utilization $\bar{U}$ of the TDD frame is obtained by weighting each state n by the probability that there is at least one active mobile not in outage:

$$\bar{U} = \sum_{n=1}^{N} \pi(n)(1 - p_0^n). \tag{31}$$

Finally, it is very important to note that the steady-state probabilities $\pi(n)$ of this model, as well as all the performance parameters, only depend on the traffic profile $(\bar{x}_{on}, \bar{t}_{off})$ through the single parameter ρ (given by relation (29)), playing a role equivalent to the "traffic intensity" of Erlang laws [15]. Indeed, the parameters $\bar{m}(n)$ that appear in relations (28) and (30) do not depend on the traffic parameters $\bar{x}_{on}$ and $\bar{t}_{off}$.

5. Throttling Policy Modeling

We now propose to adjust the generic model presented in Section 3 to take into account the throttling policy [14].

5.1. Throttling Policy

As stated in Section 2.2, a throttling policy is a scheduling policy that aims at limiting the instantaneous throughput of each active mobile to a value called $MSTR$ (*Maximum Sustained Traffic Rate*). As a result, the traffic profile of a mobile must now be characterized by three parameters: $(MSTR, \bar{x}_{on}, \bar{t}_{off})$.

The $MSTR$ regulates the maximum allowed peak rate of a connection. At each frame, the scheduler tries to allocate the right number of slots to each active mobile in order to achieve its $MSTR$. If a mobile is in outage it does not receive any slot and its throughput is degraded. If at a given time the total number of slots (N_S) is not enough to satisfy the $MSTR$ of all active users (not in outage), they all see their throughputs equally degraded. Lastly, if on the opposite there are more resources than needed, these remaining slots go unused.

5.2. Departure Rates

In order to estimate the average departure rates $\mu(n)$ corresponding to the throttling policy, we first define the following quantities.

To compensate the losses due to outage, we consider a slightly greater instantaneous bitrate than the $MSTR$, the Delivered BitRate, DBR:

$$DBR = \frac{MSTR}{1 - p_0}. \tag{32}$$

A mobile using MCS_k needs a mean number of $\bar{g}_k$ slots per frame to reach its *DBR*:

$$\bar{g}_k = \frac{DBR\, T_F}{m_k}. \tag{33}$$

Obviously, since no slots are allocated to a mobile in outage, $\bar{g}_0 = 0$.

From this, we then deduce $\bar{g}$, the average number of slots per frame needed by a mobile to obtain its $MSTR$:

$$\bar{g} = \sum_{k=1}^{K} p_k \bar{g}_k. \tag{34}$$

Knowing $\bar{g}$, we can now express the departure rates $\mu(n)$ as follows:

$$\mu(n) = \frac{N_S}{\max(n\bar{g}, N_S)}\, n\, \frac{MSTR}{\bar{x}_{on}}. \tag{35}$$

The last part of this expression ($\frac{MSTR}{\bar{x}_{on}}$) corresponds to the rate at which any of the n active mobiles completes its transfer, assuming that there are always enough available slots in the frames to satisfy the $MSTR$. The first part of this expression ($\frac{N_S}{\max(n\bar{g}, N_S)}$) represents the ratio of the global departure rate achieved by the n concurrent active transfers. Indeed, when there are n active mobiles, they need $n\bar{g}$ slots in average to obtain their $MSTR$. If $N_S \geq n\bar{g}$, they can all receive their $MSTR$ and the ratio is 1. However, if $N_S < n\bar{g}$, there are not enough resources to satisfy the demands and, as a result, the mobiles only get a portion $\frac{N_S}{n\bar{g}}$ of their $MSTR$.

Finally, note that as opposed to the full-capacity policies, we do not need to investigate the asymptotic behavior of the departure rates $\mu(n)$ since they become constant as soon as $\max(n\bar{g}, N_S) = n\bar{g}$.

5.3. Performance Parameters

By introducing the departure rates $\mu(n)$ (relation (35)) in generic expression (28), we can compute the steady-state probabilities $\pi(n)$ as:

$$\pi(n) = \frac{N!}{(N-n)!}\, \frac{\rho^n}{n! \prod_{i=1}^{n} \frac{N_S}{\max(i\bar{g}, N_S)}}\, \pi(0), \tag{36}$$

where

$$\rho = \frac{\bar{x}_{on}}{\bar{t}_{off}\, MSTR}, \tag{37}$$

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

We derive the performance parameters from the steady-state probabilities as shown in Section 3.2.2. Finally, the average utilization $\bar{U}$ of the TDD frame, is expressed as the weighted sum of the ratios between the mean number of slots needed by the n mobiles to reach their $MSTR$ and the mean number of slots they really obtain:

$$\bar{U} = \sum_{n=1}^{N} \frac{n\bar{g}}{\max(n\bar{g}, N_S)}\, \pi(n). \tag{38}$$

On a last note, let us highlight that when $\max(N\bar{g}, N_S) = N_S$, the resources of the system are always sufficient to grant a mobile its $MSTR$, even if all the N mobiles of the cell are in active transfer. As a result, by replacing $\max(N\bar{g}, N_S)$ by N_S in the expression of the departure rate $\mu(n)$ (relation (35)), we obtain that the average throughput of an active mobile (relation (30)) becomes $\bar{X} = MSTR$.

6. Multi-profile traffic Extensions

In this section, we provide non-trivial multi-profile traffic extensions for both kinds of scheduling policies (first, for the full-capacity policies, then, for the throttling policy). As a consequence, we now consider that the users are divided

into R classes of traffic, each one having a specific traffic profile $(\overline{x}^r_{on}, \overline{t}^r_{off})$, $r = 1, ..., R$ (traffic profile and traffic class are equally used in the rest of the paper). Each mobile of a given class r thus generates an infinite-length ON/OFF traffic, alternating between downloading (ON) periods characterized by an average size of $\overline{x}^r_{on}$ bits, and reading time (OFF) periods characterized by an average duration of $\overline{t}^r_{off}$ seconds. Besides, we assume that there is a fixed number N_r of mobiles belonging to each class in the cell, and that mobiles cannot change class. As a result, we now consider a total number $N = \sum_{r=1}^{R} N_r$ of mobiles with different traffic characteristics sharing the available bandwidth of the cell. Finally, we assume that all mobiles (whatever their class) have the same memoryless channel model (see Section 3.1).

6.1. Full-capacity Policy

We first propose a multi-profile traffic extension for the full-capacity models developed in Section 4 as introduced in [13].

6.1.1. Equivalent multi-class closed queueing model

We saw in Section 4.3 that, when considering a full-capacity policy, the steady-state probabilities $\pi(n)$, as well as all the performance parameters only depend on the traffic profile $(\overline{x}_{on}, \overline{t}_{off})$ through a single aggregated parameter $\rho = \frac{\overline{x}_{on}}{\overline{t}_{off}}$ (relation (29)). The key assumption of this multi-profile traffic extension is to suppose that all the performance parameters of the resulting multi-class model are still dependent of the traffic profiles through a set of aggregated parameters ρ_r given by:

$$\rho_r = \frac{\overline{x}^r_{on}}{\overline{t}^r_{off}}. \tag{39}$$

As a consequence, we can transform any class-r profile $(\overline{x}^r_{on}, \overline{t}^r_{off})$ into an equivalent profile $(\overline{x}_{on}, \overline{t'}^r_{off})$, such that $\frac{\overline{x}_{on}}{\overline{t'}^r_{off}} = \frac{\overline{x}^r_{on}}{\overline{t}^r_{off}}$. By doing so for each class, we transform the original system into an equivalent system where all classes of traffic have the same average ON size $\overline{x}_{on}$ and different average OFF durations $\overline{t'}^r_{off}$.

With this transformation, the equivalent system can be described as a mutli-class closed queueing network with two stations (see Fig. 4):

1. An IS (*Infinite Server*) station that models the mobiles in OFF periods. This station has class-dependent service rates:
$$\lambda_r = \frac{1}{\overline{t'}^r_{off}}; \tag{40}$$
2. A PS (*Processor Sharing*) station that models the active mobiles. This station has class-independent service rates $\mu(n)$ that in turn depend on the total number active mobiles (whatever their classes).

It is important to emphasize that, as all classes of the equivalent system possess the same downloading requirement (i.e., the same $\overline{x}_{on}$), the way their requests are served by the system is independent of their class, and only depends on the total number n of concurrent active mobiles and their radio conditions. Thus, the expressions of the state-dependent rates $\mu(n)$ of station 2 are exactly the same as those derived for the mono-profile traffic model in Section 4. However, this multi-profile traffic extension remains a multi-class queueing network due to the thinking times of each classes being different.

6.1.2. Performance parameters

A direct extension of the BCMP theorem [4] for stations with state-dependent rates can now be applied to this closed queueing network. The population vector is denoted by $\overrightarrow{N} = (N_1, ..., N_R)$. The detailed steady-state probabilities are expressed as follows:

$$\pi(\overrightarrow{n}) = \pi(\overrightarrow{n_1}, \overrightarrow{n_2}) = \frac{1}{G} f_1(\overrightarrow{n_1}) f_2(\overrightarrow{n_2}), \tag{41}$$

where $\overrightarrow{n_i} = (n_{i1}, ..., n_{iR})$, n_{ir} being the number of class-r mobiles present in station i,

$$f_1(\overrightarrow{n_1}) = \frac{1}{n_{11}!...n_{1R}!} \frac{1}{(\lambda_1)^{n_{11}}...(\lambda_R)^{n_{1R}}}, \tag{42}$$

$$f_2(\overrightarrow{n_2}) = \frac{n_2!}{n_{21}!...n_{2R}!} \frac{1}{\prod_{k=1}^{n_2} \mu(k)}, \tag{43}$$

n_i is the total number of mobiles (of all classes) in station i:

$$n_i = \sum_{r=1}^{R} n_{ir}, \tag{44}$$

and G, is a normalization constant:

$$G = \sum_{\overrightarrow{n_1}+\overrightarrow{n_2}=\overrightarrow{N}} f_1(\overrightarrow{n_1}) f_2(\overrightarrow{n_2}). \tag{45}$$

All the performance parameters of interest can be derived from the steady-state probabilities as follows.

We denote by $\bar{D}_r$ the average number of class-r customers departing from station 2 by unit of time, i.e., the average number of class-r mobiles completing their download by unit of time. $\bar{D}_r$ can be expressed as:

$$\bar{D}_r = \sum_{\overrightarrow{n_1}+\overrightarrow{n_2}=\overrightarrow{N}} \mu_r(\overrightarrow{n_2})\, \pi(\overrightarrow{n_1}, \overrightarrow{n_2}), \tag{46}$$

where $\mu_r(\overrightarrow{n_2})$ is the departure rate of the class-r mobiles when there are $\overrightarrow{n_2}$ active mobiles:

$$\mu_r(\overrightarrow{n_2}) = \frac{n_{2r}}{n_2} \frac{\bar{m}(n_2)\, N_S}{\bar{x}^r_{on}\, T_F}. \tag{47}$$

The average number of customers of class r in station 2, i.e., the average number of class-r active mobiles, denoted by $\bar{Q}_r$, is given by:

$$\bar{Q}_r = \sum_{\overrightarrow{n_1}+\overrightarrow{n_2}=\overrightarrow{N}} n_{2r}\, \pi(\overrightarrow{n_1}, \overrightarrow{n_2}). \tag{48}$$

The average download duration of class-r mobiles, $\bar{t}^r_{on}$, is none other than the average sojourn time of class-r customers in station 2, and is obtained from Little law:

$$\bar{t}^r_{on} = \frac{\bar{Q}_r}{\bar{D}_r}. \tag{49}$$

Knowing $\bar{t}^r_{on}$, we express the average throughput obtained by customers of class r during their transfer, denoted by $\bar{X}_r$, as:

$$\bar{X}_r = \frac{\bar{x}^r_{on}}{\bar{t}^r_{on}}. \tag{50}$$

Finally, we can compute the utilization of the TDD frame by weighting each state (where station 2 is not empty) by the probability that there is at least one mobile not in outage (and thus that the total bandwidth of the cell is used):

$$\bar{U} = \sum_{\substack{\overrightarrow{n_1}+\overrightarrow{n_2}=\overrightarrow{N} \\ \overrightarrow{n_2} \neq \overrightarrow{0}}} \pi(\overrightarrow{n_1}, \overrightarrow{n_2})\, (1 - p_0^{n_2}), \tag{51}$$

where p_0 is the outage probability.

6.2. Throttling Policy

We now propose a multi-profile traffic extension for the throttling model developed in Section 5 as presented in [14].

6.2.1. Equivalent multi-class closed queueing model

The mobiles are still divided into R classes of traffic, although this time the traffic profile of a given class-r is defined by $(MSTR_r, \overline{x}^r_{on}, \overline{t}^r_{off})$ and its associated aggregated parameter ρ_r is now:

$$\rho_r = \frac{\overline{x}^r_{on}}{\overline{t}^r_{off} MSTR_r}. \tag{52}$$

In order to apply the same idea as in the full-capacity case, we first transform any profile $(MSTR_r, \overline{x}^r_{on}, \overline{t}^r_{off})$ into an equivalent one $(MSTR, \overline{x}_{on}, \overline{t'}^r_{off})$ such that $\frac{\overline{x}_{on}}{\overline{t'}^r_{off} MSTR} = \frac{\overline{x}^r_{on}}{\overline{t}^r_{off} MSTR_r}$. After this transformation, the mobiles of the equivalent system have the same average ON size $\overline{x}_{on}$, and the same maximum instantaneous throughput $MSTR$. Thus, just like before, we can model this equivalent system as a multi-class closed queueing network with two stations: an IS station with class-dependent service rates λ_r, and a PS station with class independent but state-dependent service rates $\mu(n)$.

However, unlike for the previous extension, here we cannot directly use the same expression of the average departure $\mu(n)$ obtained in the mono-profile traffic case (relation (35)). Indeed, if we look at the expression of the steady-state probabilities derived for the mono-profile traffic model (relation (36)), we can see that they not only depend on the traffic profile through the aggregated parameter ρ, but also through the parameter $\overline{g}$ that represents the average number of slots per frame needed by a mobile to obtain its $MSTR$. We thus propose to use for $\mu(n)$ an expression very similar to relation (35):

$$\mu(n) = \frac{N_S}{\max(\overline{g}(n), N_S)} n \frac{MSTR}{\overline{x}_{on}}, \tag{53}$$

in which $MSTR$ and $\overline{x}_{on}$ are the common values of the equivalent multi-class profiles, and $\overline{g}(n)$ is the average number of slots per frame needed by n mobiles to obtain their maximum throughput.

In order to derive an expression for $\overline{g}(n)$ that takes into account the different classes of traffic, we first express DBR_r, the actual bitrate needed by a mobile of class r in order to reach its $MSTR_r$ (while compensating losses due to outage):

$$DBR_r = \frac{MSTR_r}{1 - p_0}. \tag{54}$$

We then define $\overline{g}_r$, the mean number of slots needed by a mobile of class r to obtain its $MSTR_r$:

$$\overline{g}_r = \sum_{k=1}^{K} p_k \frac{DBR_r \; T_F}{m_k}. \tag{55}$$

Second, we estimate the probabilities $\alpha_r(n)$ that an active mobile belong to class r knowing that n mobiles are active (i.e., n customers are in the PS station). These probabilities are obvious when $n = N$:

$$\alpha_r(N) = \frac{N_r}{N}, \tag{56}$$

since this means that all the $(N_1, ..., N_R)$ mobiles are active. When $n = 1$, we approximate them closely by:

$$\alpha_r(1) = \frac{N_r \rho_r}{\sum_{i=1}^{R} N_i \rho_i}. \tag{57}$$

as we know that for a given class r, the probability $\alpha_r(1)$ only increases with the number of mobiles belonging to that class, N_r, and with the intensity of the traffic they transmit, ρ_r. Knowing $\alpha_r(1)$ and $\alpha_r(N)$, we then suppose that the $\alpha_r(n)$ are a linear function of n:

$$\alpha_r(n) = an + b, \tag{58}$$

with

$$a = \frac{\alpha_r(N) - \alpha_r(1)}{N-1} \text{ and } b = \frac{N\alpha_r(1) - \alpha_r(N)}{N-1}. \quad (59)$$

Lastly, we express the average parameter $\bar{g}(n)$ as:

$$\bar{g}(n) = \sum_{r=1}^{R} n\,\alpha_r(n)\bar{g}_r. \quad (60)$$

Note that the $\alpha_r(n)$ probabilities can alternately be obtained by considering a multi-dimensional Markov chain which states $(n_1, ..., n_R)$ correspond to the detailed distribution of the current active mobiles of each class in the system. From the numerical resolution of this chain we can derive the exact values of the $\alpha_r(n)$ probabilities. We have checked on numerous examples that the exact $\alpha_r(n)$ probabilities are very well estimated by the linear approximation we propose above. In addition, the impact of this approximation is very limited as it only matters states n such that $n\bar{g}(n) < N_S$ (see relation (53)). Finally, it is important to emphasize that the use of this approximation enables to avoid the exponential complexity of solving a multi-dimensional Markov chain.

6.2.2. Performance parameters

Just like for the previous multi-profile traffic extension, we apply the extension of the BCMP theorem [4], and derive the steady-state probabilities and all the performance parameters in the exact same way. Only the expressions of the departure rates $\mu_r(\vec{n_2})$ and of the utilization of the TDD frame $\bar{U}$ must be adapted to the specificity of the throttling policy and replaced by the following ones:

$$\mu_r(\vec{n_2}) = \frac{N_S}{\max\left(\bar{g}(\vec{n_2}), N_S\right)} n_{2r} \frac{MSTR_r}{\bar{x}^r_{on}}, \quad (61)$$

and

$$\bar{U} = \sum_{\vec{n_1}+\vec{n_2}=\vec{N}} \frac{\bar{g}(\vec{n_2})}{\max\left(\bar{g}(\vec{n_2}), N_S\right)} \pi(\vec{n_1}, \vec{n_2}), \quad (62)$$

where $\bar{g}(\vec{n_2})$ represents the mean number of slots needed for the $\vec{n_2}$ active mobiles to reach their respective $MSTR_r$:

$$\bar{g}(\vec{n_2}) = \sum_{r=1}^{R} n_{2r}\bar{g}_r. \quad (63)$$

Indeed, similarly to the mono-trafic throttling case, $\bar{U}$ is expressed as the weighted sum of the ratios between the mean number of slots needed by the active mobiles to reach their $MSTR$, and the mean number of slots they really obtain.

7. Validation and Robustness

In this section, we discuss the validation of our analytical models (for mono and multi-profile traffic) through extensive simulations. We also show their robustness when confronted to more complex traffic and channel models. For this purpose has been developed a simulator that implements: i) an ON/OFF traffic generator and a wireless channel for each user; ii) a centralized scheduler allocating radio resources, i.e., slots, to active users on a frame by frame basis.

In a first phase, we validate the analytical models through simulations. In this *validation study*, the analytical models' assumptions are reproduced in the simulator. The assumptions are related to scheduling, traffic and channel models. This phase shows that describing the system by the number of active users is a sufficient approximation to obtain accurate dimensioning parameters. It also validates the analytical expressions of the average number of bits per slot $\bar{m}(n)$ for full-capacity policies and the expression of departure rates for the throttling scheme.

In a second phase, the *robustness study*, we relax the analytical models' assumptions by considering more realistic traffic and radio channel models. Through comparisons with simulation results, we show how robust the analytical model reacts towards these relaxations.

We now detail the simulation models before presenting results for both studies.

7.1. Simulation Models

7.1.1. System Parameters

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 (cf. Fig. 1). 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 considered to be 5 ms and the downlink/uplink ratio 2/3. For the sake of simplicity, we assume 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 $N_S = 450$.

7.1.2. Traffic Parameters

In our analytical models, we consider an elastic ON/OFF traffic. In the validation study, we assume that the ON data volume and OFF period are exponentially distributed as it is the case in the analytical models' assumptions. Although well adapted to reading period, the memoryless property does not always fit the reality of data traffic. For this reason, in the robustness study, we consider truncated Pareto distribution to characterize ON data volume. Recall that the mean value of the truncated Pareto distribution is given by:

$$\bar{x}_{on} = \frac{\alpha b}{\alpha - 1}\left[1 - (b/q)^{\alpha-1}\right], \tag{64}$$

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: lower and higher. These have been taken as hundred times and thousand times the mean value respectively. During simulations, the mean value in both cases (higher and lower cutoff) is the same as the exponential distribution's for the sake of comparison. The value of $\alpha = 1.2$ has been adopted from [16]. The corresponding values of parameter b for higher and lower cutoff are calculated using relation (64).

The values of parameters considered in simulations are specific to both the mono and multi-profile traffic types and are presented hereafter.

Mono-profile Traffic. Mean values of ON data volume (main page and embedded objects) and OFF period (reading time), considered in the validation study and robustness study (w.r.t. traffic distribution) for both the conventional and throttling schemes are respectively 3 Mb and 3 s. The throttling policy has one additional parameter, MSTR, which value is taken as 512 or 2048 Kbps for validation purposes. In the validation study, the behavior of the model is also observed with a higher and a lower load (i.e., with ON data volume of 1 Mb and 5 Mb). Traffic parameters for mono-profile traffic type are summarized in Tab. 1.

Multi-profile Traffic. During a simulation cycle, the total number of users, N, is partitioned among two classes (1 and 2), with equal number of users (i.e., $\frac{N}{2}$) in each class. Users in a class share the same traffic profile. Traffic parameters for full-capacity scheduling policies are summarized in Tab. 2. Three different values of N (i.e., 4, 8 and 16) are taken into account. Simulations consist of twenty cycles. Traffic profile of class 1 users is kept constant during all simulation cycles. Traffic profile of class 2 users is changed from one simulation cycle to the other. Twenty different values of $\vec{x}^2_{on}$ result into twenty different multi-profile scenarios for a given number of total users in the system i.e., one multi-profile per simulation cycle. For all these multi-profiles, value of $\vec{x}^1_{on}$ is the same i.e., 1 Mb.
Traffic parameters for the throttling scheme's validation are summarized in Tab. 3. Each class is characterized by particular values of MSTR, $\bar{x}_{on}$ and $\bar{t}_{off}$. Simulations are carried out for varying number of total users.

Table 1: Mono-profile traffic parameters.

Parameter	**Value**
Mean ON data volume $\overline{x}_{on}$	3 Mb
Mean OFF duration $\overline{t}_{off}$	3 s
$MSTR$ for throttling scheme	512/2048 Kbps
Pareto parameter α	1.2
Pareto lower cutoff q	300 Mb
Pareto higher cutoff q	3000 Mb
Pareto parameter b for lower cutoff	712926 bits
Pareto parameter b for higher cutoff	611822 bits

Table 2: Traffic parameters for multi-profile traffic and full-capacity scheduling policies.

Parameter	**Value**
Number of users in the system N	4, 8 and 16
Mean ON data volume $\overline{x}^1_{on}$ (class 1)	1 Mb
Mean ON data volume $\overline{x}^2_{on}$ (class 2)	1, 2, ..., 20 Mb
Mean OFF duration $\overline{t}^1_{off}$ (class 1)	3 s
Mean OFF duration $\overline{t}^2_{off}$ (class 2)	3 s

Table 3: Traffic parameters for multi-profile traffic and throttling scheme.

Parameter	**Value**
Mean ON data volume $\overline{x}^1_{on}$ (class 1)	3 Mb
Mean ON data volume $\overline{x}^2_{on}$ (class 2)	3 Mb
Mean OFF duration $\overline{t}^1_{off}$ (class 1)	3 s
Mean OFF duration $\overline{t}^2_{off}$ (class 2)	6 s
$MSTR_1$ (class 1)	1024 Kbps
$MSTR_2$ (class 2)	2048 Kbps

7.1.3. Channel Models

The number of bits per slot a mobile 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. Wireless channel parameters are summarized in Tab. 4. Considered MCS (including outage) and their respective number of bits transmitted per slot are given.

Table 4: Channel parameters.

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	$m_4 = 144$

A generic method for describing the channel between the base station and a mobile 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 $L\,T_F < \overline{t_{coh}}$ (the mean coherence time of the channel). In our case, and 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 (SINR is below the minimum radio quality threshold). Stationary probabilities p_k provide the long term probabilities for a mobile to receive data with MCS_k.

In our analytical study, the 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 the $(p_i)_{0\leq i\leq K}$. This corresponds to the case where $p_{ij} = p_j$ for all i. This simple approach, referred as *memoryless channel model*, is considered in the validation study, which exactly reproduces the assumptions of the analytical study. 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 of a given mobile in a frame depends on the MCS it used in the previous frame according to the FSMC presented above. The transition matrix is derived from the following equation:

$$P_T(a) = aI + (1 - a)P_T(0) \qquad 0 \leq a \leq 1, \tag{65}$$

where I is the identity matrix and a is a measure of the channel memory. A mobile indeed maintains its MCS for a certain duration with mean $\overline{t_{coh}} = 1/(1 - a)$. With $a = 0$, the transition process becomes memoryless. On the other extreme, with $a = 1$, the transition process will have infinite memory and the MCS will never change. For simulations, we have taken $a = 0.5$, so that the channel is constant in average 2 frames. This value is consistent with the coherence time given in [24] for 45 Km/h at 2.5 GHz. We call the case where all mobiles have the same channel model with memory ($a = 0.5$), *average channel model*. Note that the stationary probabilities of this model are the same as those of the memoryless model.

As the channel depends on the base station to mobile link, it is possible to refine the previous approach by considering a part of the mobiles to be under "bad" radio conditions, and the remainder under "good" ones. Bad and good radio conditions are characterized by different stationary probabilities but have the same coherence time. In the so called *combined channel model*, half of the mobiles experience good radio conditions, the other half experiences bad ones, and a is kept to 0.5 for both populations. The radio conditions are assigned to mobiles in the beginning of simulations and are not changed. For example, a mobile assigned with a bad channel state in the beginning of simulation, will keep on changing its MCS with stationary probabilities of bad radio conditions till the end of simulation.

Channel stationary probabilities for three channel models are given in Tab. 5. The respective MCS stationary probabilities for good and bad radio conditions can be obtained for example by performing system level Monte Carlo simulations and recording channel statistics close (good radio condition) or far (bad radio condition) from the base-station. For the sake of comparison, all three channel models have the same global MCS probabilities. In particular,

Table 5: Stationary probabilities for three channel models.

Channel model	Memoryless	Average	Combined	
			good 50% mobiles	bad 50% mobiles
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

those of the combined channel model are obtained by averaging stationary probabilities of good and bad radio conditions.

7.2. Validation Study

In this study, the simulator takes into account the same traffic and channel assumptions as those of analytical model. However, in the 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. Moreover, the simulator tracks the detailed status of each user, while the analysis considers aggregate states defined by the number of active users. Distributions of ON data volume and OFF period are exponential and the memoryless channel model is considered.

7.2.1. Mono-profile Traffic

Fig. 5, 6 and 7 respectively show the average channel utilization ($\bar{U}$), the average number of active users ($\bar{Q}$) and the average instantaneous throughput per user ($\bar{X}$) for the three full-capacity scheduling policies: slot sharing fairness (designated as 'Slot fair' in figures), throughput fairness ('X fair') and opportunistic ('Opp'). It is clear that simulations and analytical results are in agreement. The maximum relative error, in all cases, stays below 6% and the average relative error is less than 1%. Note that analytical results have been obtained instantaneously whereas simulations have run for several days.

Fig. 8 further proves that our analytical model is a very good description of the system: stationary probabilities $\pi(n)$ obtained by both the simulations and analysis are compared considering a cell with $N = 50$ mobiles. Again results show a perfect match between the two methods with an average relative error staying always below 9%. This means that not only average values of the output parameters can be evaluated but also higher moments with a high accuracy.

At last, Fig. 10 shows the validation for three different loads (1, 3 and 5 Mbps) and the slot fairness policy. Our model shows a comparable accuracy for all three load conditions with a maximum relative error of about 5%. The other scheduling schemes provide similar results.

The results of the validation study for throttling scheme can be found in Fig. 11, 12, 13 and 9. The maximum difference between model and simulation results in all cases is found to be less than 2%.

7.2.2. Multi-profile Traffic

The output parameters for the full-capacity scheduling policies are given in Fig. 14, 15 and 16. These parameters are plotted for twenty different multi-profile scenarios. The effect of increasing the number of users in the system is also exhibited. It is obvious from the curves depicted in the figures that the results of the analytical model are in good agreement with those of simulations. The difference between the two is less than 3% in most of the cases and less than 5% in the worst case.

If we study Fig. 15 in detail, it can be observed that $\bar{X}_1$ and $\bar{X}_2$ are not equal. We used the throughput fairness scheduling policy and the mobiles are not differentiated in any way in the PS queue. Thus, a common idea would be

that both throughputs should be the same which does not agree with the results of the figure. The difference between $\overline{X}_1$ and $\overline{X}_2$ is due to the fact that when a mobile belonging to class 1 enters the PS queue, its probability to find a given number of mobiles already present in the queue is different than the one of a mobile of class 2. As such, the mobiles of each class don't get the exact same amount of resource and hence it results in different throughputs.

Another important result that can be extracted from the figures is that our model performs equally well under low, medium and high load traffic conditions. Finally the comparison results validate the key assumption of our model, i.e., the fact that performance parameters only depend on the traffic profiles of the different classes through the aggregated parameters ρ_r given by relation (52). Indeed, if we consider the last points of all curves, it corresponds to a class 2 traffic profile of (20 Mb, 3 s) in simulations, and transformed in the analytical model into an equivalent traffic (1 Mb, 0.15 s).

The output parameters for the throttling scheme have been plotted in Fig. 17, 18 and 19. The results show that the simulations and the analytical model provide similar results not only for the overall system performance but also for each class (maximum difference is below 6%). As expected, users obtain their respective MSTR at low load and when load increases, they see their throughput proportionally degraded (Fig. 18).

7.3. Robustness Study

We now move to the robustness study, where assumptions concerning traffic and channel models made in the analytical models are relaxed in simulations. We present only the results for mono-profile traffic, the slot fairness scheduling and the throttling scheme. However, the results for multi-profile traffic and other full-capacity algorithms were tested in-house and showed similar behaviors.

In order to check the robustness of the analytical model towards the distribution of ON data volumes, simulations are carried out with a truncated Pareto distribution (with lower and higher cutoff). The results are shown in Fig. 20 for the slot fairness scheduling and in Fig. 21 for the throttling scheme. The average relative error between analytical and simulation results 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. Thus, let us take into account two different channel models such that transitions among different MCS is characterized by a process with memory: the *average channel model* and the *combined channel model*. If we look at Fig. 22 for the slot fairness policy and at Fig. 23 for the throttling scheme, it can be deduced that our analytical model shows considerable robustness even toward complex wireless channels. The average relative error is below 7% for the slot fairness scheduling and below 10% for the throttling policy. We can thus conclude that for designing a WiMAX network, channel information is almost completely included in the stationary probabilities of the MCS.

8. Dimensioning

In this section, we provide examples to demonstrate possible applications of our models while considering a mono-profile traffic scenario with the throughput fairness policy. However, results can be obtained in the same manner for any other possible configuration (i.e., any mono or multi-profile traffic scenario with any scheduling schemes) by using the according model.

8.1. Performance graphs

To obtain performance graphes, we first draw 3-dimensional surfaces where performance parameters are function of the parameters to dimension, e.g., N, the number of users in the cell and ρ, the combination of traffic parameters ($\rho = \frac{\overline{X}_{on}}{t_{off}}$, as described in Section 4.3). 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.

The average radio resource utilization of the WiMAX cell $\overline{U}$, and the average throughput per user $\overline{X}$ for any mobile in the system are presented in Fig. 24 and 25 (corresponding to the memoryless channel model presented in Table 5).

These graphs enable to directly derive any performance parameter knowing the traffic load profile, i.e., the couple (N, ρ). Each graph is the result of several thousands of input parameter sets. Obviously, any simulation tool or even any multi-dimensional Markov chain requiring numerical resolution, would have precluded the drawing of such graphs.

8.2. Dimensioning study

Here, we show how our model can be advantageously used for dimensioning issues. Two examples, each corresponding to a certain QoS criterion, are given.

In Fig. 26 we find the minimum number N_{min} of mobiles in the cell guaranteeing an average radio utilization over 50%. This kind of criterion allows operators to maximize the utilization of network resource in regard to the traffic load of their customers. To obtain the optimal value of N associated with a number of slots and a traffic load (N_S, ρ), we look for the point at the corresponding coordinates in the graph. This point is located between two level lines, and the one with the higher value gives the value of N_{min}.

The QoS criterion chosen as a second example is the throughput per user. We decided on 50 Kbps as the arbitrary value of the minimum user throughput. Now, we want to find the maximum number N_{max} of users in the cell guaranteeing this minimum throughput threshold. In Fig. 27, a given point (N_S, ρ) is located between two level lines. The line with the lower value gives N_{max}.

The graphs of Fig. 27 and 26 can be jointly used to satisfy multiple QoS criteria. For example, if we have a WiMAX cell configured to have $N_S = 450$ slots and a traffic profile given by $\rho = 300$ (e.g., $\overline{x}_{on} = 1.2$ Mb and $\overline{t}_{off} = 20$ s), Fig. 26 gives $N_{min} = 55$, and Fig. 27 gives $N_{max} = 200$. The combination of these two results recommends to have a number of users $N \in [55; 200]$ to guarantee a reasonable resource utilization and an acceptable minimum throughput to the users.

9. Conclusion

As deployment of WiMAX networks is underway, need arises for operators and manufacturers to develop dimensioning tools. In this paper, we have presented novel analytical models for elastic best effort traffic in WiMAX networks. 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, our models are remarkably straightforward. Their resolution indeed provides closed-form expressions for all the required performance parameters instantaneously. Expressions are given for four scheduling policies. Three of them are full-capacity schemes (throughput fairness, slot fairness and opportunistic scheduling). The last one, the throttling scheme, exploits the QoS parameter Maximum Sustained Traffic Rate (MSTR) foreseen by the standard to cape the maximum throughput of best effort users. Our models are also able to take into account multi-profile scenarios, in which different classes of users experience different traffic patterns. These extensions are based on original product-form queueing networks that still provide closed-form solutions for all performance parameters. Extensive simulations with various scenarios have validated the models' assumptions. The accuracy of our models 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. This shows the robust nature of our models.

Acknowledgements

The authors are thankful to Alcatel-Lucent Bell Labs France for its financial assistance in carrying out this work.

References

[1] IEEE 802.16e: IEEE 802.16e Task Group (Mobile WirelessMAN) - http://www.ieee802.org/16/tge/.
[2] IEEE Standard for local and metropolitan area networks - Part 16: Air Interface for Fixed Broadband Wireless Access Systems, 2004.
[3] 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.
[4] F. Baskett, K. Chandy, R. Muntz, and F. Palacios. Open, closed, and mixed networks of queues with different classes of customers. *Journal of the Association of Computing Machinery*, 22(2):248–260, April 1975.
[5] B. Baynat, K. Boussetta, P. Eisenmann, and N. B. Rached. A discrete-time Markovian model for GPRS/EDGE radio engineering with finite-length sessions traffic. In *Proc. of International Symposium on Performance Evaluation of Computer and Telecommunication Systems (SPECTS04)*, July 2004.
[6] B. Baynat, K. Boussetta, P. Eisenmann, and N. B. Rached. Towards an Erlang-Like formula for the performance evaluation of GPRS/EDGE networks with finite-length sessions. In *Proc. of 3rd IFIP-TC6 Networking Conference*, May 2004.
[7] B. Baynat and P. Eisenmann. Towards an Erlang-Like formula for GPRS/EDGE network engineering. In *Proc. of IEEE International Conference on Communications (ICC)*, June 2004.

[8] B. Baynat, G. Nogueira, M. Maqbool, and M. Coupechoux. An Efficient Analytical Model for the Dimensioning of WiMAX Networks. In *IFIP Networking 2009*, pages pp. 521–534, May 2009.
[9] A. Berger and Y. Kogan. Dimensioning bandwidth for elastic traffic in high-speed data networks. *IEEE/ACM Transactions on Networking*, 8(5):643–654, October 2000.
[10] T. Bonald and A. Proutiere. Wireless downlink channels: User performance and cell dimensioning. In *ACM Mobicom*, 2003.
[11] S. Borst. User-level performance of channel-aware scheduling algorithms in wireless data networks. In *IEEE Infocom*, 2003.
[12] C. Cicconetti, L. Lenzini, E. Mingozzi, and C. Eklund. Quality of Service Support in IEEE 802.16 Networks. In *IEEE Network*, 2006.
[13] S. Doirieux, B. Baynat, M. Maqbool, and M. Coupechoux. An Efficient Analytical Model for WiMAX Networks with Multiple Traffic Profiles. In *Proc. of the International Workshop on Performance and Analysis of Wireless Networks*, September 2008.
[14] 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*, June 2009.
[15] T. O. Engset. On the calculation of switches in an automatic telephone system. In *Tore Olaus Engset: The man behind the formula*, 1998.
[16] 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. In *Computer Communication Review*, October 1999.
[17] D. Heyman, T. Lakshman, and A. Neidhardt. New method for analyzing feedback protocols with applications to engineering web traffic over the internet. In *Proc. of the ACM Sigmetrics*, June 1997.
[18] G. Kulkarni, S. Adlakha, and M. Srivastava. Subcarrier Allocation and Bit Loading Algorithms for OFDMA-Based Wireless Networks. In *IEEE Trans. on Mobile Computing*, December 2005.
[19] 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. In *Proc. of VTC, p. 1231–1235*, pages 1231–1235, 2006.
[20] S. Liu and J. Virtamo. Performance Analysis of Wireless Data Systems with a Finite Population of Mobile Users. In *19th ITC*, 2005.
[21] D. Niyato and E. Hossain. A queuing-theoretic and optimization-based model for radio resource management in IEEE 802.16 broadband networks. *IEEE ToC (vol. 55)*, 2006.
[22] G. Nogueira. Modeles analytiques pour le dimensionnement des reseaux cellulaires. Phd thesis, Universite Pierre et Marie Curie, 2007. http://www-rp.lip6.fr/~nogueira/pdf/theseGN.pdf.zip.
[23] O. Grøndalen, P. Grønsund, T. Breivik and Paal Engelstad. Fixed WiMAX Field Trial Measurements and Analyses. In *Proc. of 16th IST Mobile and Wireless Communication Summit*, July 2007.
[24] K. Ramadas and R. Jain. WiMAX System Evaluation Methodology. Technical report, Wimax Forum, January 2007.
[25] D. Sivchenko, N. Bayer, B. Xu, V. Rakocevic, and J. Habermann. Internet Traffic Performance in IEEE 802.16 Networks. In *Proc. of 12th European Wireless Conference*, April 2006.
[26] A. Vinel, Y. Zhang, M. Lott, and A. Tiurlikov. Performance analysis of the random access in IEEE 802.16. In *Proc. of IEEE PIMRC, p.1596–1600*, September 2005.
[27] T. A. Yahiya, A.-L. Beylot, and G. Pujolle. Policy-based Threshold for Bandwidth Reservation in WiMAX and WiFi Wireless Networks. In *Proc. of 3rd ICWMC*, March 2007.

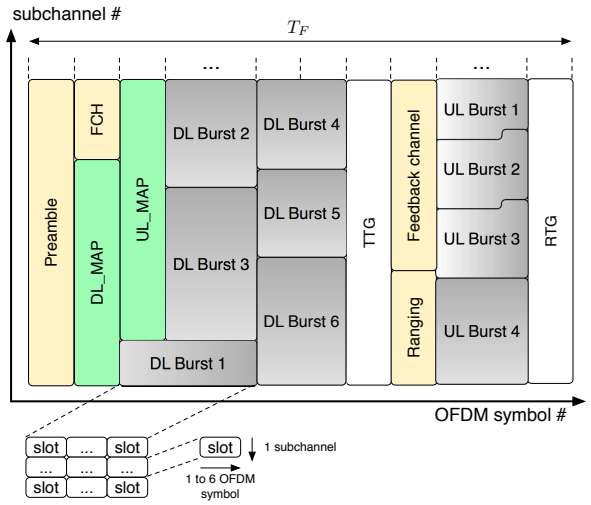

Figure 1: TDD frame structure.

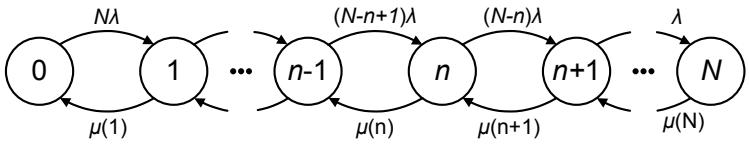

Figure 2: General CTMC with variable departure rates.

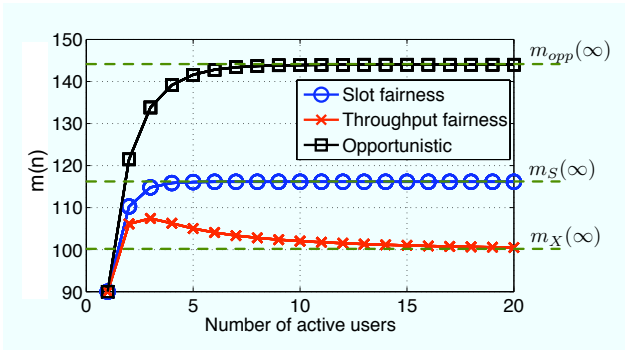

Figure 3: $\bar{m}(n)$ asymptotic behaviors.

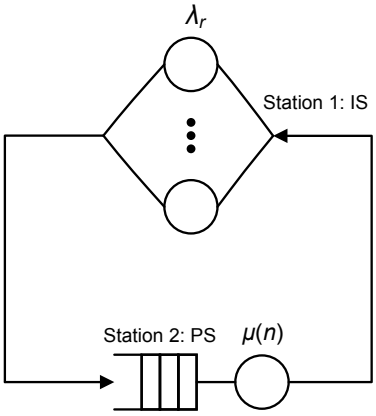

Figure 4: Closed-queueing network

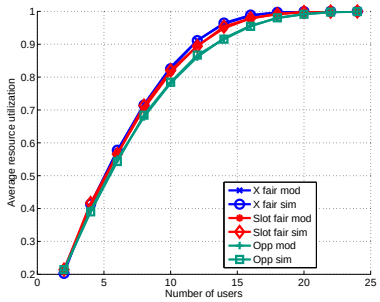

Figure 5: Average resource utilization, mono-traffic, full-capacity scheduling policies ($\overline{x}_{on}$ = 3 Mb and $\overline{t}_{off}$ = 3 s).

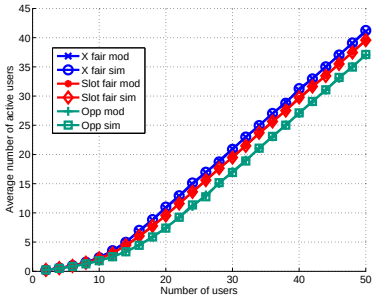

Figure 6: Average number of active users, mono-traffic, full-capacity scheduling policies ($\overline{x}_{on}$ = 3 Mb and $\overline{t}_{off}$ = 3 s).

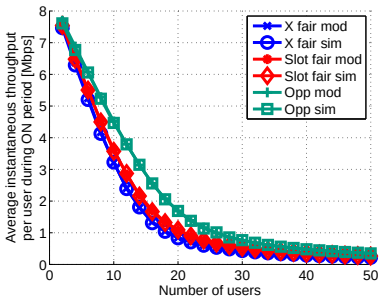

Figure 7: Average instantaneous user throughput, mono-traffic, full-capacity scheduling policies ($\overline{x}_{on}$ = 3 Mb and $\overline{t}_{off}$ = 3 s).

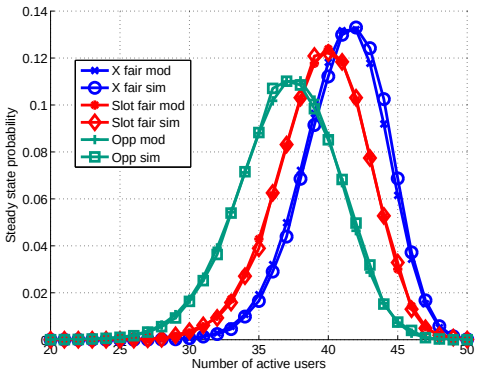

Figure 8: Steady state probabilities, mono-traffic, full-capacity scheduling policies (N = 50, $\overline{x}_{on}$ = 3 Mb and $\overline{t}_{off}$ = 3 s).

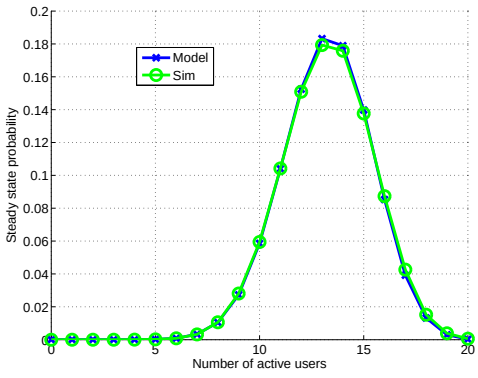

Figure 9: Steady state probabilities, mono-traffic, throttling scheme (N = 20, $\overline{x}_{on}$ = 3 Mb, $\overline{t}_{off}$ = 3 s and $MSTR$ = 512 Kbps).

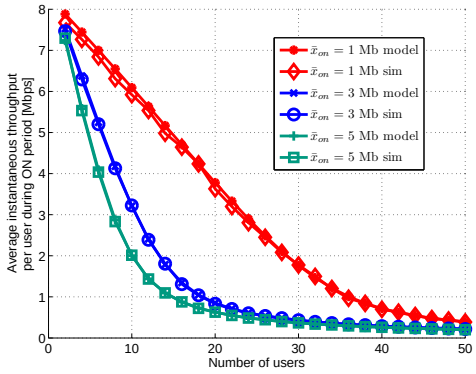

Figure 10: Average instantaneous user throughput, mono-traffic, slot fairness scheduling, different loads ($\bar{x}_{on}$ = 1, 3 and 5 Mb, $\bar{t}_{off}$ = 3 s).

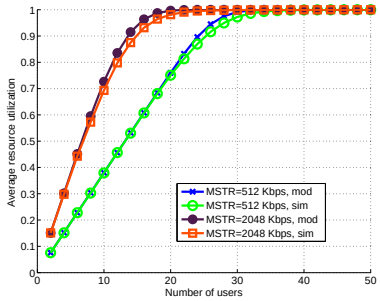

Figure 11: Average resource utilization, mono-traffic, throttling scheme ($\bar{x}_{on}$ = 3 Mb, $\bar{t}_{off}$ = 3 s, $MSTR$ = 512 and 2048 Kbps).

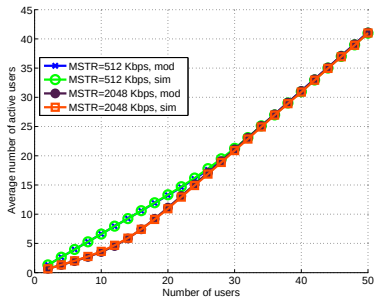

Figure 12: Average number of active users, mono-traffic, throttling scheme ($\bar{x}_{on}$ = 3 Mb, $\bar{t}_{off}$ = 3 s, $MSTR$ = 512 and 2048 Kbps).

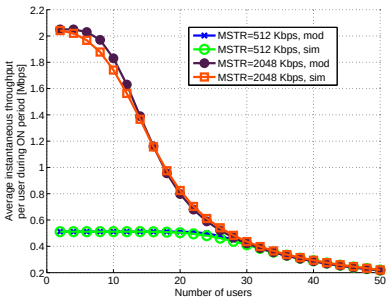

Figure 13: Average instantaneous user throughput, mono-traffic, throttling scheme, different loads ($\bar{x}_{on}$ = 3 Mb, $\bar{t}_{off}$ = 3 s, $MSTR$ = 512 and 2048 Kbps).

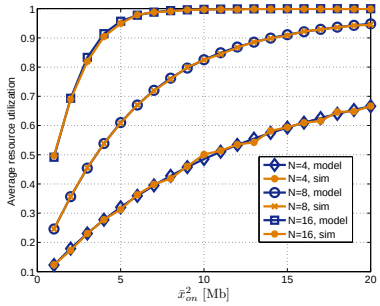

Figure 14: Average resource utilization, multi-profile, slot fairness scheduling ($\bar{x}^1_{on}$ = 1 Mb and $\bar{t}^1_{off} = \bar{t}^2_{off}$ = 3 *s*).

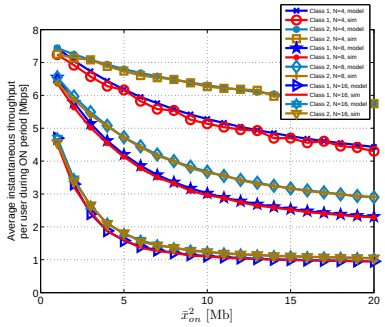

Figure 15: Average throughput per user during ON period, multi-profile, slot fairness scheduling ($\bar{x}^1_{on}$ = 1 Mb and $\bar{t}^1_{off} = \bar{t}^2_{off}$ = 3 *s*).

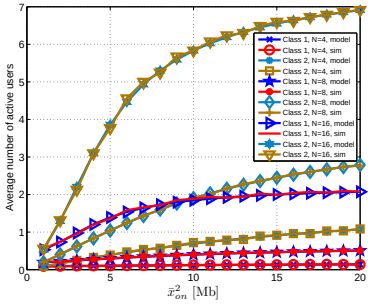

Figure 16: Average number of active users, multi-profile, slot fairness scheduling ($\bar{x}^1_{on}$ = 1 Mb and $\bar{t}^1_{off} = \bar{t}^2_{off}$ = 3 *s*).

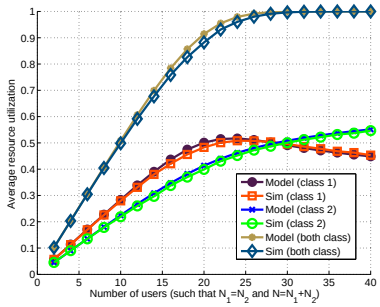

Figure 17: Average resource utilization, multi-profile, throttling scheme ($\bar{x}^1_{on}$ = $\bar{x}^2_{on}$ = 3 Mb, $\bar{t}^1_{off}$ = 3 s, $\bar{t}^2_{off}$ = 6 s, $MSTR_1$ = 1024 Kbps and $MSTR_2$ = 2048 Kbps).

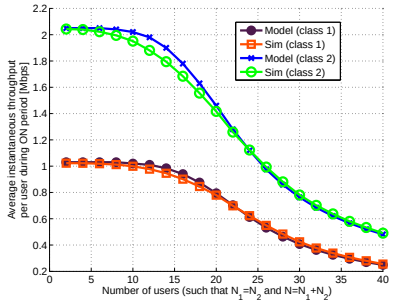

Figure 18: Average throughput per user during ON period, multi-profile, throttling scheme ($\bar{x}^1_{on}$ = $\bar{x}^2_{on}$ = 3 Mb, $\bar{t}^1_{off}$ = 3 s, $\bar{t}^2_{off}$ = 6 s, $MSTR_1$ = 1024 Kbps and $MSTR_2$ = 2048 Kbps).

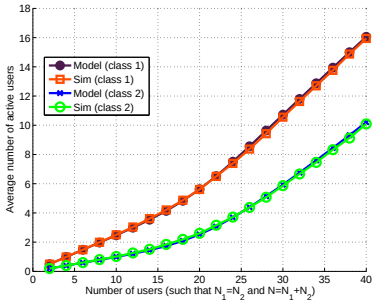

Figure 19: Average number of active users, multi-profile, throttling scheme ($\bar{x}^1_{on}$ = $\bar{x}^2_{on}$ = 3 Mb, $\bar{t}^1_{off}$ = 3 s, $\bar{t}^2_{off}$ = 6 s, $MSTR_1$ = 1024 Kbps and $MSTR_2$ = 2048 Kbps).

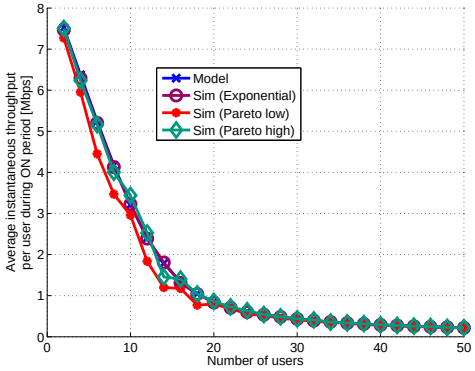

Figure 20: Average instantaneous user throughput, mono-traffic, slot fairness scheduling, different traffic distributions ($\bar{x}_{on}$ = 3 Mb and $\bar{t}_{off}$ = 3 s).

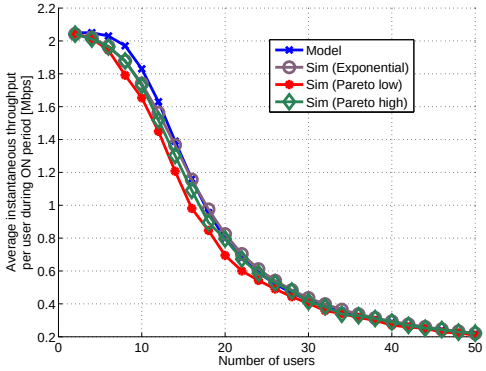

Figure 21: Average instantaneous user throughput, mono-traffic, throttling scheme schemes, different traffic distributions ($\bar{x}_{on}$ = 3 Mb, $\bar{t}_{off}$ = 3 s and $MSTR$ = 2048 Kbps).

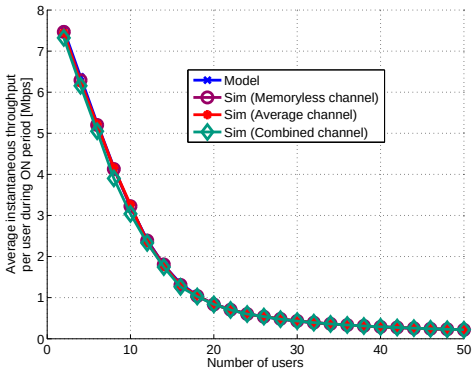

Figure 22: Average instantaneous user throughput, mono-traffic, slot fairness scheduling, different channel models ($\bar{x}_{on}$ = 3 Mb and $\bar{t}_{off}$ = 3 s).

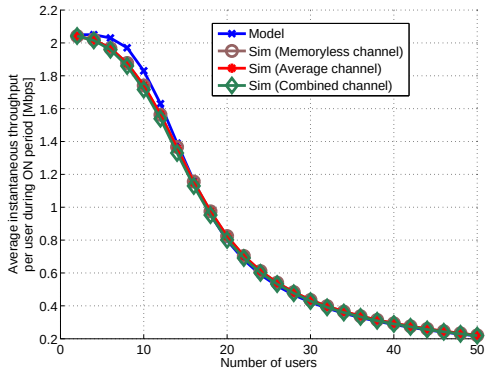

Figure 23: Average instantaneous user throughput, mono-traffic, throttling scheme, different channel models ($\bar{x}_{on}$ = 3 Mb, $\bar{t}_{off}$ = 3 s and $MSTR$ = 2048 Kbps).

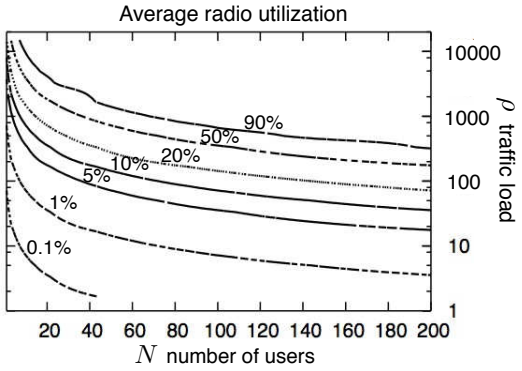

Figure 24: Average utilization $\bar{U}$.

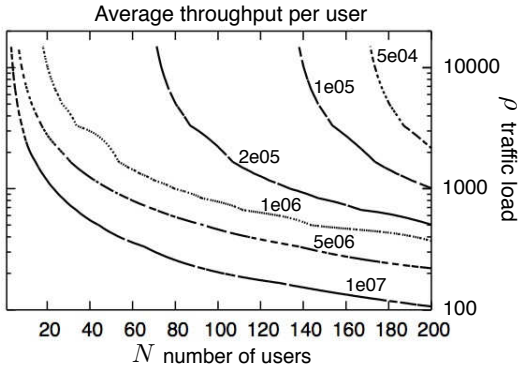

Figure 25: Average throughput per user $\bar{X}$.

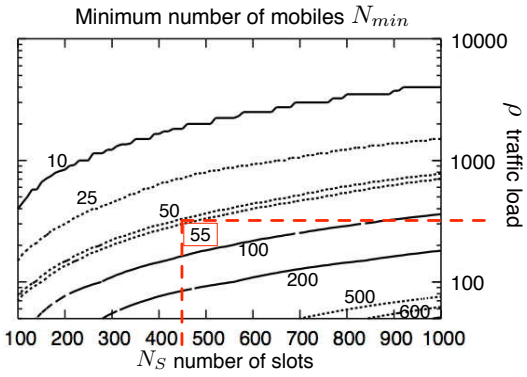

Figure 26: Dimensioning the minimum value of N guaranteeing $\bar{U} \geq 50\%$.

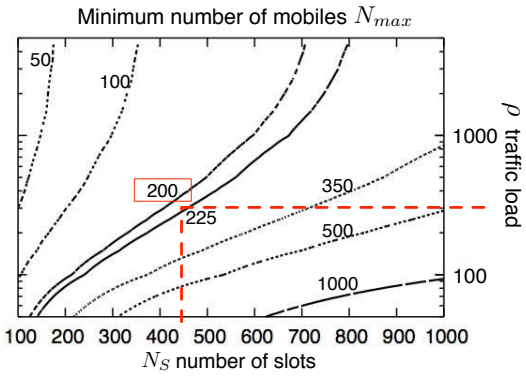

Figure 27: Dimensioning the maximum value of N guaranteeing $\bar{X} \geq 50$ Kbps.