Analytical Models for Dimensioning of OFDMA-based Cellular Networks Carrying VoIP and Best-Effort Traffic

Bruno Baynat *bruno.baynat@lip6.fr*
LIP6 - UPMC Sorbonne University - CNRS
4, place Jussieu
75005 Paris, France

Abstract

The last years have seen an exponentially growing interest for mobile telecommunication services. As a consequence, a great diversity of applications is expected to be supported by cellular networks. To answer this ever increasing demand, the ITU-R defined the requirements that the fourth generation (4G) of mobile standards must fulfill. Today, two especially promising candidates for 4G stand out: WiMAX and LTE. However, 4G cellular networks are still far from being implemented, and the high deployment costs render over-provisioning out of question. We thus propose in this paper accurate and convenient analytical models well-suited for the complex dimensioning of these promising access networks. Our main interest is WiMAX, yet, we show how our models can be easily used to consider LTE cells since both technologies are based on OFDMA. Generic Markovian models are developed specifically for three service classes defined in the WiMAX standard: UGS, ertPS and BE, respectively corresponding to VoIP, VoIP with silence suppression and best-effort traffic. First, we consider cells carrying either UGS, ertPS or BE traffic. Three methods to combine the previous models are then proposed to assume both UGS and BE traffic in the studied cell. Finally, we provide a way to easily integrate the ertPS traffic and obtain a UGS/ertPS/BE model able to account for multiple traffic profiles in each service class while keeping an instantaneous resolution. The proposed models are compared in depth with realistic simulations that show their accuracy. Lastly, we demonstrate through different examples how our models can be used to answer dimensioning issues which would be intractable with simulations.

Keywords: performance evaluation, analytical models, OFDMA, 4G, cell dimensioning, service integration.

1 INTRODUCTION

The fourth generation (4G) of mobile networks is coming to answer the ever increasing demand. Two main candidates for 4G are emerging: WiMAX (Worldwide Interoperability for Microwave Access) and 3GPP LTE (Long Term Evolution). They both propose air interfaces based on OFDMA.

WiMAX leans on the IEEE 802.16 family of standards. 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).

As for LTE, it has first been introduced in 3GPP Release 8 as a set of improvements to UMTS (Universal Mobile Telecommunications System), a widespread third generation mobile technology. An enhanced version of the LTE technology, named LTE Advanced, is under development in 3GPP Release 10 to achieve even better performance.

A great number of services such as voice, video and web are to be offered by 4G mobile networks. To this aim, several service classes have been defined in the WiMAX standard corresponding to specific QoS needs. Among them are UGS (Unsolicited Granted Service), ertPS (enhanced real-time Polling

Service) and BE (Best Effort). UGS corresponds to applications reserving a part of the resource to obtain a constant bitrate (mostly VoIP without silence suppression). ertPS has been especially designed to carry VoIP with silence suppression traffic. Finally, BE carries elastic traffic generated by web applications. Unlike WiMAX, LTE standards do not specify service classes however different mechanisms are proposed to achieve similar QoS.

Most manufacturers and operators are still under trial phases. As deployment of 4G cellular networks is under way, the need arises for fast and efficient tools used for network design and performance evaluation and able to account for these different services.

Literature on performance evaluation of cellular networks with service integration 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.

Among the latter set of papers, Borst and Hegde presented in [8] an analytical framework for wireless networks supporting a combination of streaming and elastic traffic. The authors proposed to handle the streaming connections first because of their priority over elastic traffic. Then, they used the quasi-stationary assumption first formulated by Delcoigne and al. in [12] to account for the elastic traffic without exponentially increasing the resolution complexity of their modeling. Indeed, this assumption enables to exploit the different time scales of streaming and elastic flows through astute averages to obtain the performance of the elastic flows. Note however that their study do not provide closed-form expressions of the service rates which in some cases can require high computation times or even turn intractable. In addition, this approach does not allow to observe the impact of QoS degradations on the performance parameters.

In [13], Dirani, Tarhini and Chahed designed a simple Markov chain for performance evaluation of a mobile network with one dimension corresponding to streaming and the other to elastic traffic. They too made use of the quasi-stationary assumption from [12] to simplify the resolution of their bidimensional model. However, the variations of the radio channel were not taken into account in their study. To answer this problem, Tarhini and Chahed introduced in [30] an extension of the previous model. They included in the states of the Markov chain the current channel conditions of the streaming and elastic users. They also added transitions corresponding to the probabilities that the channel conditions of a user change before the end of a connection. The resulting multi-dimensional model requires time-consuming numerical resolution, and thus, prevents in-depth dimensioning.

Niyato and Hossain presented in [27] a queuing model for bandwidth allocation in a WiMAX cell. To account for multiple services with specific QoS requirements independently from each other, they introduced a complete partitionning of the available resource among the different types of traffic using linear programming. Yet, setting fixed thresholds for each services can lead to a huge waste of bandwidth in cellular networks where radio conditions and demands of users can widely fluctuate in short amounts of time. Also, they do not account for the adaptive slot scheduling specific to OFDMA-based cellular networks.

Not specific to service integration, generic analytical models for performance evaluation of cellular networks have been proposed in [7], [6], [24]. They are mostly based on multi-class processor-sharing queues with each class corresponding to users having similar radio conditions and subsequently equal data rates. These models implicitly consider that users can only switch class between two successive data transfers. However, in broadband systems like WiMAX and LTE, radio conditions and thus data rates of a particular user can change frequently during a data transfer. In addition, the capacity of a cell may change as a result of varying radio conditions of users.

We presented in [15], [14] novel analytical models dedicated to BE traffic that take into account frame structure, precise slot sharing-based scheduling and channel quality variation of broadband wireless systems. Unlike existing models [7], [6], [24], ours are adapted to the specifics of OFDMA systems. They also offer instantaneous resolution even in multi-traffic cases: closed-form expressions are provided for all performance parameters. 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.

In this paper, we extend our models to consider UGS and ertPS traffic in addition to BE traffic. We first only consider either UGS, ertPS or BE traffic in the studied cell and provide extensions of the resulting models to take into account multi-profile traffic. Then, we propose methods to combine these models into a UGS/BE model considering a cell with both UGS and BE traffic. Finally, we integrate ertPS traffic into this model and obtain a UGS/ertPS/BE model able to account for each service class while keeping an instantaneous resolution.

To avoid confusion between WiMAX and LTE, we focus on WiMAX throughout this paper. However, we detail how the models presented here can be used for performance evaluation of an LTE cell.

The paper is organized as follows. In Section 2, the modeling assumptions are listed and the specific details on both WiMAX and LTE networks needed to understand our analytic framework are provided. Sections 3, 4 and 5 present our analytical models for UGS, ertPS and BE service classes respectively. As a first step, methods to combine these models into an UGS/BE model are introduced and compared in Section 6. Then, in Section 7, we integrate ertPS traffic into our model and validate the resulting UGS/ertPS/BE model through comparisons with simulations. Lastly, Section 8, provides examples of WiMAX dimensioning processes using this model.

2 MODELING ASSUMPTIONS

Our analytical models stand on several assumptions related to the system, the channel and the traffic. The validity of these assumptions has been thoroughly discussed in our study of BE traffic [15]. To avoid any possible confusion, between WiMAX and LTE specifics, we initially only consider the WiMAX technology. The assumptions shared by the models mentioned in this paper are presented and, wherever required, related particulars of WiMAX systems are specified. In addition, various notations are introduced. Lastly, we explain how these assumptions can be adapted to consider an LTE cell.

2.1 WiMAX Modeling

2.1.1 System

A WiMAX time division duplex (TDD) frame is divided in slots using Orthogonal Frequency Division Multiplexing (OFDM). A slot occupies space both in the time and frequency domains and is the smallest unit of resource that can be allocated to a mobile. A frame is comprised of two parts: one is dedicated to uplink and the other to downlink. Besides, a portion of the frame is used for overhead (e.g., UL_MAP and DL_MAP). The duration T_F of this TDD frame is equal to 5 ms [3].

1. We consider a single WiMAX cell and focus in this paper on the downlink part which is a critical portion of asymmetric data traffic. However, our models can also be used for the dimensioning of the uplink part in a similar way.
2. We assume that there is a mean number of slots available for data transmission in the downlink part of each TDD frame denoted by $\bar{N}_S$. This number is a mean value because the size of the downlink part can vary with the downlink to uplink bandwidth ratio which can be adjusted dynamically over time. The size of the overhead, increasing with the number of multiplexed transmissions per frame, can also affect the number of available downlink slots. However, we consider that the small variations of the amount of overhead are not significant in regard to the size of the downlink part.
3. In the case of UGS (respectively ertPS) traffic, we consider that there is a limit V_{max} (resp. W_{max}) to the number of simultaneous calls accepted in the cell. On the contrary, regarding BE traffic, we assume that all mobiles can simultaneously be in active transfer. As a consequence, any BE connection demand will be accepted and no blocking can occur.

Note that this last assumption implicitly states that no admission control of BE connections is implemented in the system. However, the BE model can be easily modified to account for a system with an

admission controller limiting the number of simultaneous active BE transfers. Indeed, as detailed in [15], we just have to truncate the BE Markov chain accordingly.

2.1.2 Channel

One of the important features of IEEE 802.16e is link adaptation: different modulation and coding schemes (MCS) allows a dynamic adaptation of the transmission to the radio conditions. Many subcarrier permutations defining how the pilot and data subcarriers should be distributed over the subchannels are proposed in the standard. As the number of data subcarriers per slot is the same for all permutation schemes [11], the number of bits carried by a slot for a given MCS is constant. The selection of appropriate MCS is carried out according to the value of signal to interference plus noise ratio (SINR). In case of outage, i.e., if the SINR is too low, no data can be transmitted without error. We denote the radio channel states as: MCS_k, $1 \leq k \leq K$, where K is the number of MCS. By extension, MCS_0 represents the outage state. The number of bits transmitted per slot by a mobile using MCS_k is denoted by m_k. For the particular case of outage, $m_0 = 0$.

The radio link quality in broadband wireless networks like WIMAX is highly variable. As such, the MCS used by a given mobile can change very often.

4. We assume that each mobile sends a feedback channel estimation on a frame by frame basis, and thus, the base station can change its MCS 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. Table 1 presents examples of MCS and their associated probabilities.

MCS	bits per slot	probability
Outage	$m_0 = 0$	$p_0 = 0.02$
QPSK-1/2	$m_1 = 48$	$p_1 = 0.12$
QPSK-3/4	$m_2 = 72$	$p_2 = 0.31$
16QAM-1/2	$m_3 = 96$	$p_3 = 0.08$
16QAM-3/4	$m_4 = 144$	$p_4 = 0.47$

Table 1: Channel Parameters.

As a result, our analytical model only depends upon stationary probabilities of using the different MCS and does not explicitly take into account the radio channel dynamics. These probabilities can be accurately obtained from famous statistical fading models such as Rayleigh or Rician models as shown in numerous publications including [20], [25], [31]. In addition, other methods can be considered. By example, in [19] a spatial model is used while in [26] a semi-analytical approach is proposed based on an interpolation of simulation results.

Finally, note that the robustness of this assumption to temporal channel correlation has been validated through extensive simulations considering radio channels with memory as shown in section 7.2.

2.1.3 Traffic

The traffic model is based on the following assumptions.

5. We assume that there is a fixed number N of mobiles sharing the available bandwidth of the cell. The numbers of mobiles generating UGS, ertPS or BE traffic present in the cell are denoted, respectively, by N_{ugs}, N_{ertps} and N_{be}.

Note that operators find finite population models more suitable for the dimensioning of a cell. Indeed, they have means to estimate the number of users they will have to serve in a cell and, as such, consider

those models more appropriate. However, our models can be easily adapted to Poisson arrivals should an infinite population assumption be considered [15].

6. Each of the N mobiles is assumed to generate an infinite length ON/OFF traffic. In the case of UGS and ertPS traffics, an ON period corresponds to a call and is characterized by its duration. In the case of BE traffic, an ON period corresponds to the download of an element (e.g., a web page including all embedded objects). As opposed to UGS and ertPS ON periods, the downloading duration depends on the system load and the radio link quality, so BE ON periods must be characterized by their size. Lastly, in each case, an OFF period corresponds to an idle time independent of the system load and, as such, is characterized by its duration.

7. We assume that UGS and ertPS ON durations, BE ON sizes and each OFF duration are exponentially distributed. We denote by $\bar{t}_{on}^{ugs}$ and $\bar{t}_{on}^{ertps}$ the average durations of UGS and ertPS ON periods (in seconds), by $\bar{x}_{on}^{be}$ the average size of ON data volumes (in bits) and by $\bar{t}_{off}^{ugs}$, $\bar{t}_{off}^{ertps}$ or $\bar{t}_{off}^{be}$ the average durations of OFF periods (in seconds).

Memoryless BE traffic distributions are strong assumptions that have been validated by numerous theoretical results. Several works on insensitivity (e.g., [5], [7], [18]) 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. Moreover, note that we compare in Section 7.2 the results of our analytical model with those of simulations considering a truncated Pareto ON size distribution. These comparisons tend to prove that insensitivity still holds or is at least a very good approximation. Thus, memoryless distributions are the most obvious choice to model BE traffic.

8. We consider absolute priorities between each service class. As such, at each frame, the available slots are first allocated to UGS, then to ertPS and at last to BE connections.

No specific inter-class scheduling is suggested in the WiMAX standard. However, the QoS requirements of each service class suggest that UGS traffic should always be served first, followed by ertPS traffic and finally by BE traffic [23], [28], [29].

2.2 WiMAX to LTE Modeling

We detail here how a 3GPP LTE cell can be easily considered instead of a WiMAX cell as both technologies are based on OFDMA. In addition, we introduce the LTE particulars required to apprehend these modifications whenever needed.

2.2.1 System

LTE frames are organized differently than WiMAX frames. Each LTE frame lasts 10 ms and is divided into 10 subframes [1]. Contrary to WiMAX, a new scheduling is not done at each frame but at each of these subframes. As a consequence, we need, in our modeling, to consider different time intervals between two consecutive schedulings:

- WiMAX: frame duration, $T_F = 5$ ms.
- LTE: subframe duration, $T_F = 1$ ms.

An LTE subframe comprises of resource blocks which, similarly to WiMAX slots, occupy space in both time and frequency domains. Although both LTE and WiMAX use OFDMA as their multiple access scheme, they specify different smallest data allocation units. Indeed, the smallest data unit that can be allocated to an LTE user is formed by a pair of resource blocks (i.e., two consecutive resource blocks in either the frequency or the time domain). So, $\bar{N}_S$, the mean quantity of downlink resource available during one time interval must be adjusted as follows:

- WiMAX: $\bar{N}_S$ is the mean number of downlink slots in a frame.
- LTE: $\bar{N}_S$ is the mean number of downlink resource block pairs in a subframe.

2.2.2 Channel

WiMAX and LTE systems share the same adaptive modulation and coding mechanism so the MCS_k and p_k parameters stay the same. But, the smallest data allocation unit differs for each technology and, as such, the m_k now represent different values whether we consider WiMAX or LTE:

- WiMAX: numbers of transmitted bits per slot using MCS_k.
- LTE: numbers of transmitted bits per resource block pair using MCS_k,

$$m_k = 2\, N_{re}\, m_{re\,k}, \tag{1}$$

where N_{re} is the number of downlink resource elements in a resource block, which depends on the antenna settings, and $m_{re\,k}$ is the number of transmitted bits per resource element using MCS_k.

2.2.3 Traffic

The traffic generated by the mobiles present in the cell is not affected by the considered mobile technology. Thus, no adjustments are required to our traffic modeling assumptions.

Lastly, unlike WiMAX standards, LTE does not specify service classes. The traffic separation is possible instead by defining Evolved Packet System bearers (bearers for short) which provide differential treatment for traffic with differing QoS requirements. The QoS parameters of the bearers enable to consider the same scheduling assumption than we do for our WiMAX models. As such, we keep those assumptions for our modeling of LTE.

3 UGS MODEL

The UGS (*Unsolicited Grant Service*) service class has been designed to support real-time applications periodically generating fixed-size data packets (e.g., VoIP without silence suppression). In this section, we provide the models we use to characterize the UGS traffic of a WiMAX cell. Let us highlight that even though they are based on the famous Engset model [16], we still need to adapt to the specifics of OFDMA and our channel model. Indeed, it would not be possible to formulate crucial performance parameters such as the mean throughput achieved by the mobiles and the mean utilization of the frames otherwise.

We consider in these models that only UGS mobiles (i.e., mobiles generating only UGS traffic) are present in the cell we study. How to model cells with traffic of several service classes is addressed in Sections 6 and 7.

3.1 Mono-Traffic Model

An UGS call corresponds to an utilization of the resource in circuit mode. The reserved bit rate associated to each UGS connection is called Guaranteed Bit Rate (GBR). In a first phase, no distinctions between users are made: all mobiles are considered statistically identical. As such, we assume that the N_{ugs} users are generating infinite-length ON/OFF constant bit rate traffics with the same traffic profile $(GBR, \bar{t}_{on}^{ugs}, \bar{t}_{off}^{ugs})$.

The amount of resource, i.e., the number of slots, needed at each frame by an UGS connection to achieve its GBR varies with the MCS it uses. In order to prevent the losses caused by outage periods, we assume that an UGS connection reserves a slightly greater bit rate than its GBR, called Delivered Bit Rate (DBR):

$$DBR = \frac{GBR}{1 - p_0}. \tag{2}$$

We model this system by a continuous-time Markov chain (CTMC) where each state v, represents the total number of concurrent UGS calls, regardless of the coding scheme they use. The maximum number of UGS calls accepted being $V_{ugs} = \min(N_{ugs}, V_{max})$, this chain is thus made of $V_{ugs} + 1$ states as shown in Fig 1.

- A transition out of a generic state v to state $v+1$ occurs when a mobile in OFF period starts its call. This "arrival" transition corresponds to one mobile among the $(N_{ugs} - v)$ in OFF period, ending its idle time, and is performed with a rate $(N_{ugs} - v)\lambda_{ugs}$, where λ_{ugs} is defined as the inverse of the idle time between two calls: $\lambda_{ugs} = \frac{1}{t_{off}^{ugs}}$.
- On the opposite, a transition out of a generic state v to state $v - 1$ occurs when a mobile in ON period finishes its call. This "departure" transition corresponds to one mobile among the v in ON period, ending its call, and is performed with a rate $v\mu_{ugs}$, where μ_{ugs} is defined as the inverse of the average ON duration: $\mu_{ugs} = \frac{1}{t_{on}^{ugs}}$.

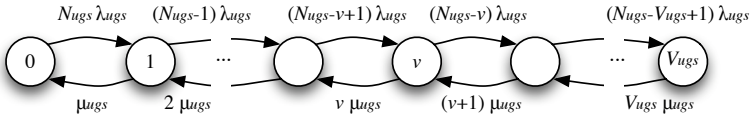


Figure 1: Mono-traffic UGS CTMC.

This results in the famous Engset model [16] which steady state probabilities $\pi_{ugs}(v)$ of having v current calls are derived as:

$$\pi_{ugs}(v) = \frac{\rho_{ugs}^{v}}{v!} \frac{N_{ugs}!}{(N_{ugs} - v)!} \pi_{ugs}(0), \tag{3}$$

with

$$\rho_{ugs} = \frac{\lambda_{ugs}}{\mu_{ugs}}. \tag{4}$$

and $\pi_{ugs}(0)$ obtained by normalization.

We then deduce the performance parameters as follows. First, the probability of rejecting a call P_{rej} is expressed:

$$P_{rej} = \frac{\pi_{ugs}(V_{ugs})(N_{ugs} - V_{ugs})}{\sum_{v=0}^{V_{ugs}} \pi_{ugs}(v)(N_{ugs} - v)}. \tag{5}$$

We compute $\bar{Q}_{ugs}$, the mean number of current UGS calls as:

$$\bar{Q}_{ugs} = \sum_{v=1}^{V_{ugs}} v\pi_{ugs}(v). \tag{6}$$

To attain its DBR, a mobile using MCS_k needs g_k slots:

$$g_k = \frac{DBR\, T_F}{m_k}. \tag{7}$$

Obviously, no slots are allocated to a mobile in outage so $g_0 = 0$. The available resource being limited, an UGS mobile does not always achieve its GBR if V_{max} is too big. Thus, we also derive $\bar{X}_{ugs}$, the instantaneous throughput obtained by an UGS mobile:

$$\bar{X}_{ugs} = \sum_{v=1}^{V_{ugs}} \frac{\pi_{ugs}(v)}{1 - \pi_{ugs}(0)} \sum_{\substack{(v_0, ..., v_K) = (0, ..., 0) | \\ v_0 + ... + v_K = v \\ v_0 \neq v}}^{(v,...,v)} p(v_0, ..., v_K) \frac{\bar{N}_S}{\max\left(\sum_{k=1}^{K} v_k g_k, \bar{N}_S\right)} GBR, \tag{8}$$

where $p(v_0, ..., v_K)$ is the probability that the v mobiles are distributed among the K MCS as $(v_0, ..., v_K)$ (v_k being the number of mobiles using MCS_k):

$$p(v_0, ..., v_K) = \binom{v}{v_0, ..., v_K} \left(\prod_{k=0}^{K} p_k^{v_k} \right), \tag{9}$$

with $\binom{v}{v_0,...,v_K}$ the multinomial coefficient, and $\left(\frac{\bar{N}_S}{\max\left(\sum_{k=1}^{K} v_k g_k, \bar{N}_S\right)} GBR \right)$ corresponds to the throughput achieved by a mobile when the v connections are distributed in $(v_0, ..., v_K)$. So, to obtain $\bar{X}_{ugs}$ we average this throughput for every possible distributions and remove the case when no UGS connections are active (i.e., state $v = 0$). Note that when there is no degradation, this expression leads to $\bar{X}_{ugs} = GBR$.

Finally, $\bar{U}_{ugs}$, the average utilization of the TDD frame by UGS connections is expressed as:

$$\bar{U}_{ugs} = \sum_{v=1}^{V_{ugs}} \pi_{ugs}(v) \sum_{\substack{(v_0, ..., v_K) = (0, ..., 0)| \\ v_0 + ... + v_K = v \\ v_0 \neq v}}^{(v,...,v)} p(v_0, ..., v_K) \frac{\sum_{k=1}^{K} v_k g_k}{\max\left(\sum_{k=1}^{K} v_k g_k, \bar{N}_S\right)}. \tag{10}$$

Note that when V_{ugs} is small enough to guarantee that UGS calls are never degraded, i.e., $V_{ugs} \leq \frac{\bar{N}_S}{g_1}$, this expression can be greatly simplified:

$$\bar{U}_{ugs} = \sum_{v=1}^{V_{ugs}} \frac{\bar{g}(v)}{\bar{N}_S} \pi_{ugs}(v), \tag{11}$$

where $\bar{g}(v)$ is the mean number of slots needed by v UGS calls to obtain their DBR:

$$\bar{g}(v) = v \sum_{k=1}^{K} p_k g_k. \tag{12}$$

3.2 Multi-Traffic Extension

We now relax the assumption that all users have the same traffic profile. To do so, we distribute the mobiles among R traffic profiles defined by $(GBR_r, \bar{t}_{on}^r, \bar{t}_{off}^r)$. Thus, the mobiles of a given profile r generate an infinite-length ON/OFF traffic, with a guaranteed bit rate of GBR_r bits per second, an average ON duration of $\bar{t}_{on}^r$ seconds and an average OFF duration of $\bar{t}_{off}^r$ seconds. We consider that there is a fixed number N_{ugs}^r of mobiles belonging to each profile in the cell. So, there are $N_{ugs} = \sum_{r=1}^{R} N_{ugs}^r$ users in the cell with different traffic profiles.

Note that for the sake of clarity, the *ugs* indexes are removed from the $\bar{t}_{on}^r$ and $\bar{t}_{off}^r$ notations in this section.

Similarly to the mono-traffic model, we define DBR_r, the bit rate demanded by a class-r mobile:

$$DBR_r = \frac{GBR_r}{1 - p_0}. \tag{13}$$

To model this system, we use the multi-class extension of the Engset model [22]. The associated CTMC contains as much dimensions as considered traffic profiles and each of its states is characterized by a specific R-tuple $(v_1, ..., v_R)$ where v_r is the number of active connections of class r.

- A transition out of a generic state $(v_1, ..., v_r, ..., v_R)$ to state $(v_1, ..., v_r + 1, ..., v_R)$ occurs when a class-r mobile in OFF period starts its call. This "arrival" transition is performed with a rate $\lambda_r(v_1, ..., v_R) = (N_r - v_r)\lambda_r$, where λ_r is defined as the inverse of a class-r average idle time: $\lambda_r = \frac{1}{\bar{t}_{off}^r}$.

- A transition out of a generic state $(v_1, ..., v_r, ..., v_R)$ to state $(v_1, ..., v_r - 1, ..., v_R)$ occurs when a class-r mobile in ON period ends its call. This "departure" transition is performed with a rate $\mu_r(v_1, ..., v_R) = v_r \mu_r$, where μ_r is defined as the inverse of a class-r average call duration: $\mu_r = \frac{1}{t_{on}^r}$.

We also assume that V_{max}, the limit on the maximum number of concurrent UGS calls, is observed regardless of the classes they belong to. So, we define V_{ugs}^r, the maximum number of possible class-r simultaneous calls as:

$$V_{ugs}^r = \min\left(N_{ugs}^r, V_{max}\right). \tag{14}$$

However, nothing prevents from considering more complex limitations (e.g., privileging certain traffic profiles over others). Indeed, we then only need to adapt the possible states of the CTMC and the resolution of the model stays the same [21].

Fig. 2 presents this model's CTMC when considering only two traffic profiles ($R = 2$) and $V_{max} \leq \min(N_{ugs}^1, N_{ugs}^2)$.

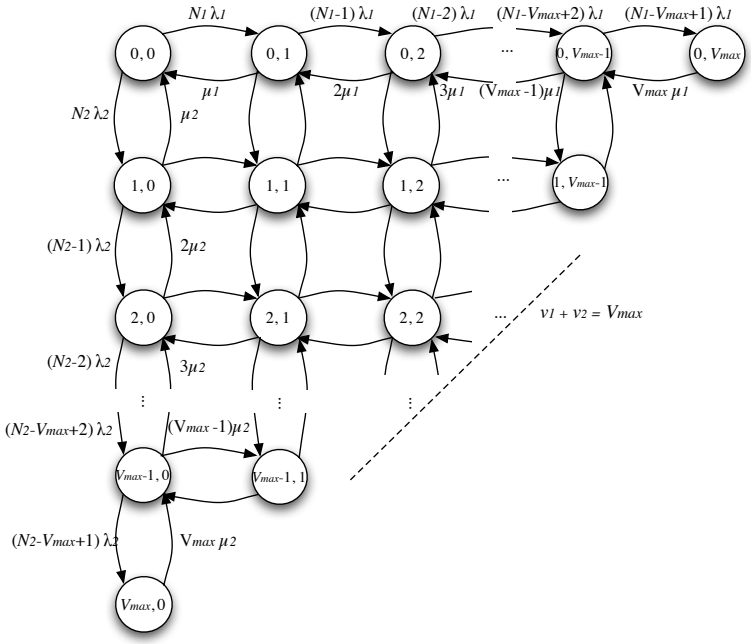


Figure 2: 2-dimensional multi-traffic UGS CTMC.

The steady state probabilities $\pi_{ugs}(v_1, ..., v_R)$ of having $(v_1, ..., v_R)$ concurrent UGS calls are given by:

$$\pi_{ugs}(v_1, ..., v_R) = \left(\prod_{r=1}^{R} \frac{\rho_r^{v_r}}{v_r!} \frac{N_{ugs}^r!}{(N_{ugs}^r - v_r)!} \right) \pi_{ugs}(0, ..., 0), \tag{15}$$

with

$$\rho_r = \frac{\lambda_r}{\mu_r}, \tag{16}$$

and $\pi_{ugs}(0, ..., 0)$ obtained by normalization.

The performance parameters are derived from the steady-state probabilities as follows. The probability

of rejecting a class-r call P^r_{rej} is given by:

$$P^r_{rej} = \frac{\sum\limits_{\substack{(v_1,...,v_R)=(0,...,0)|\\ v_1+...+v_R=V_{max}}}^{(V^1_{ugs},...,V^R_{ugs})} \pi_{ugs}(v_1,...,v_R)(N^r_{ugs}-v_r)}{\sum\limits_{\substack{(v_1,...,v_R)=(0,...,0)|\\ v_1+...+v_R\leq V_{max}}}^{(V^1_{ugs},...,V^R_{ugs})} \pi_{ugs}(v_1,...,v_R)(N^r_{ugs}-v_r)}. \quad (17)$$

Obviously, this probability is null when $N_{ugs} \leq V_{max}$ as there can be no blocking in this case.

We can compute $\bar{Q}^r_{ugs}$, the mean number of concurrent UGS calls belonging to class r as:

$$\bar{Q}^r_{ugs} = \sum_{\substack{(v_1,...,v_R)=(0,...,0)|\\ v_1+...+v_R\leq V_{max}}}^{(V^1_{ugs},...,V^R_{ugs})} v_r\ \pi_{ugs}(v_1,...,v_R). \quad (18)$$

Finally, by first defining $\bar{g}(v_1,...,v_R)$, the mean number of slots needed by $(v_1,...,v_R)$ UGS mobiles to achieve their respective GBR_r:

$$\bar{g}(v_1,...,v_R) = \sum_{r=1}^{R} v_r \sum_{k=1}^{K} p_k \frac{DBR_r\ T_F}{m_k}, \quad (19)$$

we can then express $\bar{X}^r_{ugs}$, the instantaneous throughput achieved by class-r mobiles as:

$$\bar{X}^r_{ugs} = \sum_{\substack{(v_1,...,v_R)=(0,...,0)|\\ v_1+...+v_R\leq V_{max}}}^{(V^1_{ugs},...,V^R_{ugs})} \frac{\pi_{ugs}(v_1,...,v_R)}{1-p_{v_r=0}} \frac{\bar{N}_S}{\max\left(\bar{g}(v_1,...,v_R),\bar{N}_S\right)} GBR_r, \quad (20)$$

with $p_{v_r=0}$ the probability that no class-r mobile is active:

$$p_{v_r=0} = \sum_{\substack{(v_1,...,v_R)=(0,...,0)|\\ v_1+...+v_R\leq V_{max}\\ v_r=0}}^{(V^1_{ugs},...,V^R_{ugs})} \pi_{ugs}(v_1,...,v_R), \quad (21)$$

and compute $\bar{U}_{ugs}$, the average utilization of the TDD frame as:

$$\bar{U}_{ugs} = \sum_{\substack{(v_1,...,v_R)=(0,...,0)|\\ v_1+...+v_R\leq V_{max}}}^{(V^1_{ugs},...,V^R_{ugs})} \frac{\bar{g}(v_1,...,v_R)}{\max\left(\bar{g}(v_1,...,v_R),\bar{N}_S\right)} \pi_{ugs}(v_1,...,v_R). \quad (22)$$

4 ERTPS MODEL

The ertPS (*enhanced real-time Polling Service*) service class has been especially added to the WiMAX standards in order to carry traffic from VoIP with silence suppression. As such, an ertPS call only occupies the resource during the talk spurts of the conversation. We show, in this section, how our UGS model can be easily adapted to take into account the impact of silence suppression on the cell capacity. To this aim, we now consider a cell carrying only ertPS traffic.

A telephonic conversation can be seen as a succession of ON periods (talk spurts) and OFF periods (silences). As shown in [9], [10], the durations of these periods can be accurately modeled by exponential

distributions with mean values $\bar{t}_{talk}$ et $\bar{t}_{sil}$ respectively. Note that [10] recommends to set these values to $\bar{t}_{talk} = 1.2$ s et $\bar{t}_{sil} = 1.8$ s.

Knowing this, we can account for the effect of silence suppression in a mono-traffic scenario by adjusting the expression of g_k, the number of slots needed by a mobile using MCS_k to attain its DBR, as:

$$g_k = \frac{\bar{t}_{talk}}{\bar{t}_{talk} + \bar{t}_{sil}} \frac{DBR\, T_F}{m_k}. \tag{23}$$

Similarly, in a multi-traffic scenario, we just need to modify the expression of $\bar{g}(v_1, ..., v_R)$, the mean number of slots needed by $(v_1, ..., v_R)$ mobiles using to achieve their respective GBR_r, as:

$$\bar{g}(v_1, ..., v_R) = \sum_{r=1}^{R} v_r \sum_{k=1}^{K} p_k \frac{\bar{t}_{talk}}{\bar{t}_{talk} + \bar{t}_{sil}} \frac{DBR_r\, T_F}{m_k}, \tag{24}$$

to consider the bandwidth saved by the silence suppression.

The rest of our UGS modeling approach still stands whether we consider UGS or ertPS traffic. So, the computing of the ertPS steady state probabilities and performance parameters is the same as in the UGS model.

5 BE MODELS

The BE (*Best Effort*) service class of the WiMAX standard has been planned to carry the traffic of applications without QoS guarantee (e.g., web applications). In this section, we provide an overview of the analytical models we use to perform the traffic analysis of this service class. In all the models presented here, we consider that only BE mobiles (i.e., mobiles generating only BE traffic) are present in the studied cell. Models for cells carrying traffic belonging to several service classes are detailed in Sections 6 and 7.

Our main concern in this paper is to show how these BE models can be used as stepping stones for the study of cells carrying traffics of several service classes. As a consequence, we only introduce the various parameters needed for the dimensioning procedure and do not detail their expressions. However, note that these models have already been fully explained, discussed and validated in [15], [14].

5.1 Scheduling

Several scheduling schemes can be considered. In [15], we focused on three traditional schemes:

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

Lastly, in [14], we proposed an alternative scheduling called throttling which forces an upper bound on the users' throughputs, the Maximum Sustained Traffic Rate (MSTR):

- The throttling scheduling tries to allocate at each frame 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 available slots is not enough to satisfy the MSTR of all active users (not in outage), they all see their throughputs equally degraded.

5.2 Mono-Traffic Model

As a first step, we do not make any distinction between users and consider all mobiles as statistically identical. Thus, we consider that the N_{be} users are generating infinite-length ON/OFF BE traffics with the same traffic profile $(\bar{x}_{on}^{be}, \bar{t}_{off}^{be})$.

We model this system by a continuous-time Markov chain (CTMC) where each state n, represents the total number of concurrent active BE connections, regardless of the coding scheme they use. So, the resulting CTMC is made of $N_{be}+1$ states as depicted in Fig 3.

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

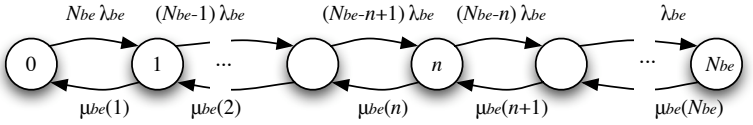


Figure 3: Mono-traffic BE CTMC with state-dependent departure rates.

Obviously, the main difficulty of the model resides in estimating the aggregate departure rates $\mu_{be}(n)$.

If we consider either the instantaneous throughput fairness, the slot sharing fairness or the opportunistic policy, they are expressed as follows [15]:

$$\mu_{be}(n) = \frac{\bar{m}(n)\,\bar{N}_S}{\bar{x}_{on}^{be}\,T_F}, \tag{25}$$

where $\bar{m}(n)$ is the average number of bits transmitted per slot when there are n concurrent active transfers. These parameters are strongly dependent on the scheduling policy. As a consequence, we provide their different expressions for each policy.

With the slot sharing policy:

$$\bar{m}(n) = \sum_{\substack{(n_0,\ldots,n_K)=(0,\ldots,0)|\\ n_0+\ldots+n_K=n\\ n_0\neq n}}^{(n,\ldots,n)} \frac{n!}{n-n_0}\sum_{k=1}^{K} m_k n_k \prod_{k=0}^{K}\frac{p_k^{n_k}}{n_k!}. \tag{26}$$

With the instantaneous throughput fairness policy:

$$\bar{m}(n) = \sum_{\substack{(n_0,\ldots,n_K)=(0,\ldots,0)|\\ n_0+\ldots+n_K=n\\ n_0\neq n}}^{(n,\ldots,n)} \frac{(n-n_0)\,n!\prod_{k=0}^{K}\frac{p_k^{n_k}}{n_k!}}{\sum_{k=1}^{K}\frac{n_k}{m_k}}. \tag{27}$$

With the opportunistic policy:

$$\bar{m}(n) = \sum_{k=1}^{K} m_k \left(1 - \sum_{j=k+1}^{K} p_j\right)^n \left(1 - \left(1 - \frac{p_k}{\sum_{j=0}^{k} p_j}\right)^n\right). \tag{28}$$

If we consider the throttling policy, the departure rates $\mu_{be}(n)$ become:

$$\mu_{be}(n) = \frac{\bar{N}_S}{\max\left(n\bar{g}, \bar{N}_S\right)} n \frac{MSTR}{\bar{x}_{on}^{be}}. \tag{29}$$

with $\bar{g}$, the average number of slots per frame needed by a mobile to obtain its MSTR:

$$\bar{g} = T_F \, MSTR \sum_{k=1}^{K} \frac{p_k}{(1-p_0)m_k}. \tag{30}$$

Once the departure rates $\mu_{be}(n)$ have been determined, the steady state probabilities $\pi_{be}(n)$ of having n concurrent transfers in the cell, can easily be derived from the birth-and-death structure of the Markov chain:

$$\pi_{be}(n) = \left(\prod_{i=1}^{n} \frac{(N_{be} - i + 1)\lambda_{be}}{\mu_{be}(i)}\right) \pi_{be}(0), \tag{31}$$

where $\pi_{be}(0)$ is obtained by normalization.

Note that the $\bar{x}_{on}^{be}$ and $\bar{t}_{off}^{be}$ traffic parameters are only involved in this last expression through their ratio. As a consequence, we define the intensity ρ_{be} of the traffic generated by a mobile:

$$\rho_{be} = \frac{\bar{x}_{on}^{be}}{\bar{t}_{off}^{be}}. \tag{32}$$

The following performance parameters of the system can be obtained from the steady state probabilities. The average number of active users $\bar{Q}_{be}$ is expressed as:

$$\bar{Q}_{be} = \sum_{n=1}^{N_{be}} n \, \pi_{be}(n), \tag{33}$$

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

$$\bar{D}_{be} = \sum_{n=1}^{N_{be}} \mu_{be}(n) \, \pi_{be}(n). \tag{34}$$

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

$$\bar{t}_{on}^{be} = \frac{\bar{Q}_{be}}{\bar{D}_{be}}. \tag{35}$$

and compute the average throughput $\bar{X}_{be}$ obtained by one BE mobile in active transfer as:

$$\bar{X}_{be} = \frac{\bar{x}_{on}^{be}}{\bar{t}_{on}^{be}}. \tag{36}$$

Finally, we can express the average utilization $\bar{U}_{be}$ of the TDD frame. This last parameter depends on the scheduling policy. Indeed, with the instantaneous throughput fairness, the slot sharing fairness or

the opportunistic policy, the cell is considered fully utilized as long as there is at least one active mobile not in outage:

$$\bar{U}_{be} = \sum_{n=1}^{N_{be}} (1 - p_0^n)\pi_{be}(n). \tag{37}$$

However, if we consider the throttling policy, $\bar{U}_{be}$ is now 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 obtain:

$$\bar{U}_{be} = \sum_{n=1}^{N_{be}} \frac{n\bar{g}}{\max\left(n\bar{g}, \bar{N}_S\right)} \pi_{be}(n). \tag{38}$$

5.3 Multi-Traffic Extension

Now, we relax the assumption that all users have the same traffic profile. To this aim, we associate to each mobile one of the R traffic profiles, $(\bar{x}_{on}^r, \bar{t}_{off}^r)$. The mobiles of a given profile r thus generate an infinite-length ON/OFF traffic, with an average ON size of $\bar{x}_{on}^r$ bits and an average reading time of $\bar{t}_{off}^r$ seconds. We assume that there is a fixed number N_{be}^r of mobiles belonging to each profile in the cell. As a consequence, there are $N_{be} = \sum_{r=1}^{R} N_{be}^r$ users in the cell with different traffic profiles.

Similarly to the notations in Section 3.2, the *be* indexes are removed from the $\bar{x}_{on}^r$, $\bar{t}_{on}^r$ and $\bar{t}_{off}^r$ notations in this section.

To compute the performance parameters, we first transform this system into an equivalent one where all profiles of traffic have the same average ON size $\bar{x}_{on}$, and different average OFF durations $\bar{t'}_{off}^r$, such that [15]:

$$\frac{\bar{x}_{on}}{\bar{t'}_{off}^r} = \frac{\bar{x}_{on}^r}{\bar{t}_{off}^r}. \tag{39}$$

With this transformation, the equivalent system can be described as a mutli-class closed queuing network with two stations as shown by Fig. 4:

1. An *Infinite Servers* (IS) station that models mobiles in OFF periods. This station has profile-dependent service rates $\lambda_r = \frac{1}{\bar{t'}_{off}^r}$;
2. A *Processor Sharing* (PS) station that models active mobiles. This station has profile-independent service rates $\mu_{be}(n)$ that in turn depend on the total number active mobiles (whatever their profiles). They are given by the same relations than the departure rates of the mono-traffic model (see relations 25 and 29)

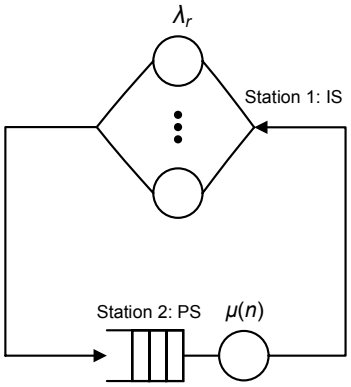


Figure 4: Closed-queueuing network.

A direct extension of the BCMP theorem [4] for stations with state-dependent rates can now be applied to this closed queueing network. The detailed steady state probabilities are expressed as follows:

$$\pi_{be}(\overrightarrow{n_1},\overrightarrow{n_2}) = \frac{1}{G} f_1(\overrightarrow{n_1}) f_2(\overrightarrow{n_2}), \tag{40}$$

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

$$f_1(\overrightarrow{n_1}) = \frac{1}{n_{11}!...n_{1R}!} \frac{1}{(\lambda_1)^{n_{11}}...(\lambda_R)^{n_{1R}}} \tag{41}$$

$$f_2(\overrightarrow{n_2}) = \frac{(n_{21}+...+n_{2R})!}{n_{21}!...n_{2R}!} \frac{1}{\prod_{k=1}^{n_2} \mu_{be}(k)}, \tag{42}$$

and G is the normalization constant:

$$G = \sum_{\overrightarrow{n_1}+\overrightarrow{n_2}=\overrightarrow{N}_{be}} f_1(\overrightarrow{n_1}) f_2(\overrightarrow{n_2}). \tag{43}$$

All the performance parameters of interest can be derived from the steady state probabilities as follows. The average number of profile-r active mobiles, $\bar{Q}_r$, is given by:

$$\bar{Q}_r = \sum_{\overrightarrow{n_1}+\overrightarrow{n_2}=\overrightarrow{N}_{be}} n_{2r}\, \pi_{be}(\overrightarrow{n_1},\overrightarrow{n_2}), \tag{44}$$

and the average number of profile-r mobiles completing their download by unit of time, $\bar{D}_r$, can be expressed as:

$$\bar{D}_r = \sum_{\overrightarrow{n_1}+\overrightarrow{n_2}=\overrightarrow{N}_{be}} \mu(n_2)\, \pi_{be}(\overrightarrow{n_1},\overrightarrow{n_2}), \tag{45}$$

with $n_2 = \sum_{r=1}^{R} n_{2r}$.

The average download duration of profile-r mobiles, $\bar{t}^r_{on}$, is obtained from Little's law:

$$\bar{t}^r_{on} = \frac{\bar{Q}_r}{\bar{D}_r}. \tag{46}$$

And we can then calculate the average throughput obtained by customers of profile r during their transfer, denoted by $\bar{X}_r$, as:

$$\bar{X}_r = \frac{\bar{x}^r_{on}}{\bar{t}^r_{on}} \tag{47}$$

Finally, the utilization $\bar{U}_{be}$ of the TDD frame is expressed differently whether we consider the instantaneous throughput fairness, the slot sharing fairness, the opportunistic policy:

$$\bar{U}_{be} = \sum_{\overrightarrow{n_1}+\overrightarrow{n_2}=\overrightarrow{N}_{be}} (1-p_0^{n_2})\pi_{be}(\overrightarrow{n_1},\overrightarrow{n_2}), \tag{48}$$

or the throttling policy:

$$\bar{U}_{be} = \sum_{\overrightarrow{n_1}+\overrightarrow{n_2}=\overrightarrow{N}_{be}} \frac{n_2\bar{g}}{max\left(n_2\bar{g}, \bar{N}_S\right)} \pi_{be}(\overrightarrow{n_1},\overrightarrow{n_2}). \tag{49}$$

Again, fully detailed explanations on the multi-traffic model and the different relations are available in [14]. Finally, let us add that [14] also includes a method to consider the throttling policy and traffic profiles with different MSTR values.

6 UGS/BE MODEL

Using the previous models, we can consider a cell carrying either UGS or BE traffic. However, we now desire to take into account a cell with both kinds of traffic without increasing the resolution complexity of the resulting model.

To do so, we first propose three methods to combine both mono-traffic UGS and BE models. Then, we compare results brought by these methods and conclude on the best one to use. Lastly, we explain how to extend the resulting UGS/BE model to multiple traffic profiles in both service classes.

6.1 Combining the Mono-Traffic Models

As a first step, we only consider one UGS traffic profile and one BE traffic profile.

6.1.1 General approach

The WiMAX standard does not recommend any scheduling between service classes. However, it is common sens and widely admitted in the literature that UGS calls preempt BE traffic [23], [28], [29]. Indeed, UGS connections reserve a part of the resource in each frame and BE mobiles share whatever is left to them. This has two major consequences: i) our UGS model is sufficient to characterize the UGS traffic since UGS calls are not affected by the presence of the BE traffic; ii) the performance evaluation of the BE mobiles are strongly dependent on the portion of resource left to them by the UGS traffic at each frame. As such, the part of the resource, i.e., the mean number of slots, left to the BE connections needs to be evaluated and accounted for in the characterization of the BE traffic.

Our general approach to combine both UGS and BE models consists in the following steps:

1. We compute the steady state probabilities $\pi_{ugs}(v)$ of having v active UGS connections by only using the UGS model.
2. We then estimate the mean number of slots occupied by these v UGS calls and deduce the mean number of slots available to BE connections when there are v active UGS calls.
3. For each possible value of v, we solve a CTMC of the BE model with the $\bar{N}_S$ parameter set to the corresponding mean number of slots available. We obtain the steady state probabilities $\pi_{be}(n|v)$ of having n active BE connections provided that there are v concurrent UGS calls.
4. We express the steady state probabilities $\pi(v,n)$ of having simultaneously both v active UGS and n active BE connections as:
$$\pi(v,n) = \pi_{ugs}(v)\pi_{be}(n|v). \tag{50}$$

Three methods to combine the UGS and BE models are proposed, each one corresponding to a different level of precision in evaluating the portion of the resource left to the BE mobiles.

6.1.2 DTL (detailed) method

As its name indicates, this method is the most precise of the three. Indeed, for each possible value v of simultaneous active UGS connections, we specifically consider each possible distributions $(v_0, ..., v_K)$ of the v connections among the $K+1$ MCS (including outage). And, for each of these distributions, we compute $\bar{N}_S^{be}(v_0, ..., v_K)$, the corresponding mean number of slots left to the BE connections:

$$\bar{N}_S^{be}(v_0, ..., v_K) = \bar{N}_S - \bar{N}_S^{ugs}(v_0, ..., v_K), \tag{51}$$

with

$$\bar{N}_S^{ugs}(v_0, ..., v_K) = min\left(\sum_{k=1}^{K} v_k g_k, \bar{N}_S\right). \tag{52}$$

Then, for each $\bar{N}_S^{be}(v_0, ..., v_K)$, we solve a corresponding BE CTMC and obtain the steady state probabilities $\pi_{be}(n|(v_0, ..., v_K))$ of having n active BE transfers provided that there are $(v_0, ..., v_K)$ active UGS connections.

From these, we first deduce the probabilities $\pi_{be}(n|v)$:

$$\pi_{be}(n|v) = \sum_{\substack{(v_0,...,v_K)=(0,...,0)| \\ v_0+...+v_K=v}}^{(v,...,v)} p(v_0, ..., v_K).\pi_{be}(n|(v_0, ..., v_K)), \tag{53}$$

then, the probabilities $\pi(v, n)$ with relation 50.

Note that for the distributions where no slots remain for BE traffic ($\bar{N}_S^{be}(v_0, ..., v_K) = 0$), we set:

$$\pi_{be}(n|(v_0, ..., v_K)) = \begin{cases} 1 & \text{if } n = N_{be} \\ 0 & \text{else} \end{cases} \tag{54}$$

because we consider that BE connections keep on being initiated but none of them can end without available resource.

Lastly, let us highlight that this method, illustrated by Fig.5(a), requires to solve as much BE CTMC as there are possible values of $\bar{N}_S^{be}(v_0, ..., v_K)$.

6.1.3 AVG (averaged) method

This method has been proposed to significantly reduce the number of BE CTMC to solve. We now only evaluate $\bar{N}_S^{ugs}(v)$, the mean numbers of slots occupied by v UGS calls by averaging the $\bar{N}_S^{ugs}(v_0, ..., v_K)$ as follows:

$$\bar{N}_S^{ugs}(v) = \sum_{\substack{(v_0,...,v_K)=(0,...,0)| \\ v_0+...+v_K=v \\ v_0 \neq v}}^{(v,...,v)} p(v_0, ..., v_K)\bar{N}_S^{ugs}(v_0, ..., v_K). \tag{55}$$

So, this time, to obtain the $\pi_{be}(n|v)$ probabilities, we only solve one CTMC for each possible value of v, considering that $\bar{N}_S^{be}(v) = \bar{N}_S - \bar{N}_S^{ugs}(v)$ slots remain available to BE transfers when there are v active UGS connections. Thus, with this method (see Fig. 5(b)), the number of BE CTMC to solve is reduced to $V_{ugs} + 1$.

Let us highlight that this method is not based on the quasi-stationary assumption introduced in [12] by Delcoigne and al. Indeed, here, we do not need to make any assumption on the time scales of UGS and BE traffics.

6.1.4 AGG (aggregated) method

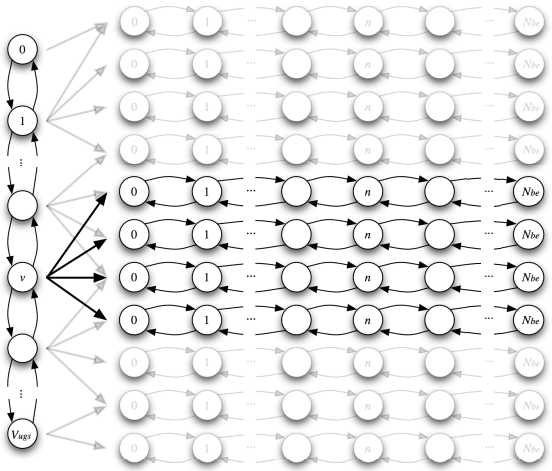
This last method, as shown in Fig. 5(c), is the most straightforward of the three: only one BE CTMC has to be solved. We first compute $\bar{N}_S^{ugs}$, the mean number of slots occupied by UGS connections as:

$$\bar{N}_S^{ugs} = \sum_{v=1}^{V_{ugs}} \pi_{ugs}(v)\bar{N}_S^{ugs}(v). \tag{56}$$

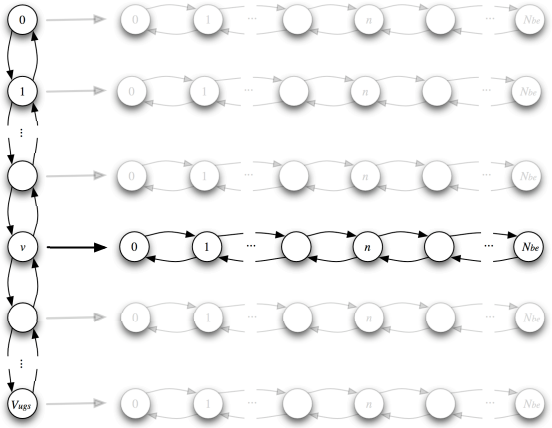
Then we only have to solve the BE CTMC corresponding to $\bar{N}_S^{be} = \bar{N}_S - \bar{N}_S^{ugs}$ slots available to the BE mobiles to obtain the probabilities $\pi_{be}(n)$ and the various performance parameters.

6.1.5 Performance parameters

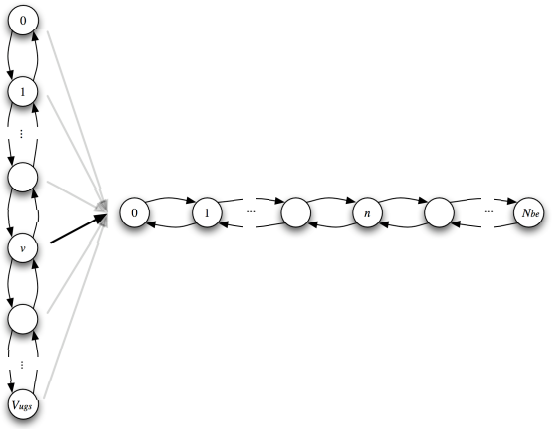
UGS



(a) DTL method



(b) AVG method



(c) AGG method

Figure 5: Three methods to combine the UGS and BE models. (UGS CTMC is vertical, BE ones horizontal.)

As stated earlier, the UGS traffic is not affected by the presence of the BE traffic. So, the UGS performance parameters are computed as detailed in Section 3.1.

BE DTL/AVG

When using either the DTL or AVG method, the BE performance parameters are derived from the steady state probabilities $\pi(v,n)$ as follows:

$\bar{Q}_{be}$, the mean number of active BE users is given by:

$$\bar{Q}_{be} = \sum_{n=1}^{N_{be}} n \sum_{v=0}^{V_{ugs}} \pi(v,n). \tag{57}$$

$\bar{D}_{be}$, the mean number of BE departures per unit of time, depends on the number of slots left to the BE mobiles. As such, its expression differs whether we consider the DTL or AGG method. So, for DTL, we consider all the possible distributions of the v UGS connections among the K MCS:

$$\bar{D}_{be} = \sum_{v=0}^{V_{ugs}} \sum_{\substack{(v_0,...,v_K)=(0,...,0)| \\ v_0+...+v_K=v}}^{(v,...,v)} p(v_0,...,v_K) \sum_{n=1}^{N_{be}} \bar{N}_S^{be}(v_0,...,v_K)\mu_{be}(n)\pi(v,n), \tag{58}$$

whereas for AVG we only use the $\bar{N}_S^{be}(v)$:

$$\bar{D}_{be} = \sum_{v=0}^{V_{ugs}} \sum_{n=1}^{N_{be}} \bar{N}_S^{be}(v)\mu_{be}(n)\pi(v,n). \tag{59}$$

In both cases, we obtain from Little's law $\bar{t}_{on}^{be}$, the average duration of a BE transfer:

$$\bar{t}_{on}^{be} = \frac{\bar{Q}_{be}}{\bar{D}_{be}}, \tag{60}$$

and deduce $\bar{X}_{be}$, the average throughput achieved by a BE mobile:

$$\bar{X}_{be} = \frac{\bar{x}_{on}^{be}}{\bar{t}_{on}^{be}}. \tag{61}$$

Finally, the average utilization of the TDD frame by BE transfers, $\bar{U}_{be}$, needs to be adapted to both methods but also to the considered BE scheduling. With DTL and the slot sharing, the instantaneous throughput fairness or the opportunistic policy:

$$\bar{U}_{be} = \sum_{v=0}^{V_{ugs}} \sum_{\substack{(v_0,...,v_K)=(0,...,0)| \\ v_0+...+v_K=v}}^{(v,...,v)} p(v_0,...,v_K) \frac{\bar{N}_S^{be}(v_0,...,v_K)}{\bar{N}_S} \sum_{n=1}^{N_{be}} (1-p_0^n)\pi(v,n). \tag{62}$$

With DTL and the throttling policy:

$$\bar{U}_{be} = \sum_{v=0}^{V_{ugs}} \sum_{\substack{(v_0,...,v_K)=(0,...,0)| \\ v_0+...+v_K=v}}^{(v,...,v)} p(v_0,...,v_K). \frac{\bar{N}_S^{be}(v_0,...,v_K)}{\bar{N}_S} \sum_{n=1}^{N_{be}} \frac{n\bar{g}}{\max\left(n\bar{g}, \bar{N}_S^{be}(v_0,...,v_K)\right)} \pi(v,n). \tag{63}$$

With AVG and the slot sharing, the instantaneous throughput fairness or the opportunistic policy:

$$\bar{U}_{be} = \sum_{v=0}^{V_{ugs}} \frac{\bar{N}_S^{be}(v)}{\bar{N}_S} \sum_{n=1}^{N_{be}} (1-p_0^n)\pi(v,n). \tag{64}$$

With AVG and the throttling policy:

$$\bar{U}_{be} = \sum_{v=0}^{V_{ugs}} \frac{\bar{N}_S^{be}(v)}{\bar{N}_S} \sum_{n=1}^{N_{be}} \frac{n\bar{g}}{\max\left(n\bar{g}, \bar{N}_S^{be}(v)\right)} \pi(v,n). \tag{65}$$

BE AGG

As for the AGG method, the BE performance parameters are obtained using the $\pi_{be}(n)$ as explained in Section 5.2. We just have to replace $\bar{N}_S$ by $\bar{N}_S^{be}$ in the various expressions.

Only the average utilization of the resource by BE traffic, $\bar{U}_{be}$, needs to be adapted following the considered BE scheduling. With the slot sharing, the instantaneous throughput fairness or the opportunistic policy:

$$\bar{U}_{be} = \frac{\bar{N}_S^{be}}{\bar{N}_S} \sum_{n=1}^{N_{be}} (1 - p_0^n) \pi_{be}(n). \tag{66}$$

With the throttling policy:

$$\bar{U}_{be} = \frac{\bar{N}_S^{be}}{\bar{N}_S} \sum_{n=1}^{N_{be}} \frac{n\bar{g}}{\max\left(n\bar{g}, \bar{N}_S^{be}\right)} \pi_{be}(n). \tag{67}$$

6.2 Comparison

Here, we compare results obtained with the DTL, AVG and AGG methods. The channel, cell and traffic parameters are summarized in Tables 1 and 2.

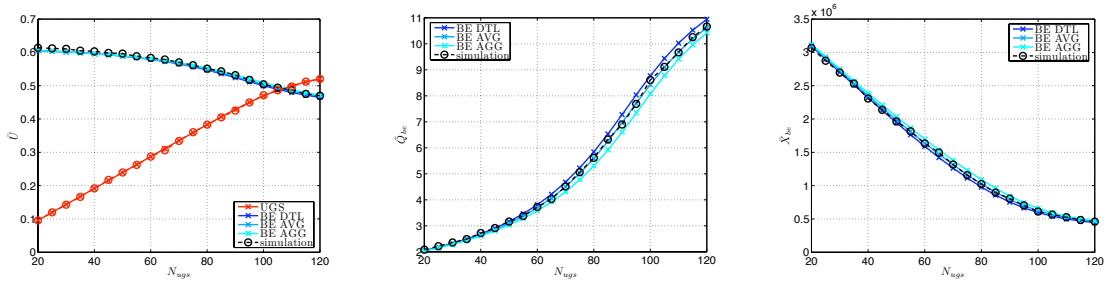
Parameter		Value
Number of slots per frame, N_S		450
Frame duration, T_F		5 ms
Limit on UGS calls, V_{max}		40
BE scheduling		slot fairness
UGS	Number of UGS mobiles, N_{ugs}	20 to 120
	Guaranteed bit rate, GBR	128 Kbps
	Mean ON duration, $\bar{t}_{on}^{ugs}$	60 s
	Mean OFF duration, $\bar{t}_{off}^{ugs}$	120 s
BE	Number of BE mobiles, N_{be}	40
	Mean ON size, $\bar{x}_{on}^{be}$	1 Mbits
	Mean OFF duration, $\bar{t}_{off}^{be}$	6 s

Table 2: Cell and traffic parameters.

The duration T_F of one TDD frame of WiMAX is 5 ms. We consider $\bar{N}_S = 450$ slots per frame available for downlink. This value roughly corresponds to a system bandwidth of 10 MHz, a downlink/uplink ratio of 2/3, a Fast Fourrier Transform size equal to 2048, a PUSC subcarrier permutation and an average protocol overhead length of 4 symbols.

We observe the behaviors of $\bar{U}$, $\bar{Q}_{be}$ and $\bar{X}_{be}$ parameters given by each method while considering a cell with an increasing number of UGS mobiles, N_{ugs} and a fixed number of BE mobiles N_{be}. The maximum number of concurrent UGS calls accepted in the cell is set to $V_{max} = 40$ and BE connections are scheduled using the slot fairness policy.

Fig. 6(a) presents the evolution of the average utilization of the TDD frame by UGS and BE traffics when the UGS traffic load increases. Obviously, the portions of the frames occupied by UGS connections increases with the number of UGS mobiles in the cell and the BE transfers get less and less resource.



(a) Average utilization, $\bar{U}$ of the TDD frame by UGS and BE connections.

(b) Average number of concurrent BE transfers, $\bar{Q}_{be}$.

(c) Instantaneous throughput of a BE mobile, $\bar{X}_{be}$.

Figure 6: Comparison of the UGS/BE methods: Customary BE traffic parameters when UGS traffic increases.

So, as shown in Fig. 6(b) and 6(c) respectively, BE mobiles stay in ON periods longer and achieve smaller throughputs.

Moreover, these figures also provide simulation results. Indeed, as a first step to validate our approach, we have developed a simulator that integrates an ON/OFF traffic generator, a wireless channel for each user and a centralized scheduler allocating radio resources, i.e., slots, to active users on a frame by frame basis. The simulator allocates the resources to active users at each frame serving first the UGS connections and then the BE connections. At the beginning of a frame, a new MCS is drawn for each active mobile according to the channel probabilities p_k. Then, the scheduler allocates slots to those mobiles depending on their MCS and their service class. In the simulations presented here, a mobile generates an exponentially distributed ON/OFF traffic and is subject to a memoryless channel. Note however that results of more realistic simulations are provided in section 7.2 to show the robustness of our analytical modeling toward the channel model and the distribution of BE ON sizes.

In all the figures, we can see that the results obtained with each of the three UGS/BE methods and the simulations are always very close. Furthermore, we have compared these methods and the simulations in all sorts of configurations (different traffic loads, different bit rates reserved by UGS calls, different BE schedulings, etc.). Each time, the results they gave proved to be almost equivalent: the difference between them remained below 3%.

This leads us to conclude that as straightforward as the AGG method may appear, its precision is sufficient to efficiently combine the UGS and BE models.

6.3 Multi-Traffic Extensions

As stated above, the 3 methods give similar results so we now only consider the AGG method. The simplicity of this method enables us to extend the resulting UGS/BE model to multiple traffic profiles in both service classes as follows.

6.3.1 Multi-trafic UGS

Again, UGS traffic preempts BE traffic. So, to consider multiple UGS traffic profiles, we just use the multi-traffic extension introduced in Section 3.2. The only modification to the UGS/BE model resides in the expression of $\bar{N}_S^{ugs}$:

$$\bar{N}_S^{ugs} = \sum_{\substack{(v_1,...,v_R)=(0,...,0)| \\ v_1+...+v_R \leq V_{max}}}^{(V_{ugs}^1,...,V_{ugs}^R)} \pi_{ugs}(v_1,...,v_R) \sum_{v=1}^{V_{ugs}} \min\left(\bar{N}_S^{ugs}(v_1,...,v_R), \bar{N}_S\right), \quad (68)$$

where $\bar{N}_S^{ugs}(v_1,...,v_R)$ is the mean number of slots occupied by UGS connections distributed in $(v_1,...,v_R)$ among the R traffic profiles:

$$\bar{N}_S^{ugs}(v_1,...,v_R) = \sum_{r=1}^{R} \sum_{\substack{(j_0,...,j_K)=(0,...,0)| \\ j_0+...+j_K=v_r}}^{(v_r,...,v_r)} p(j_0,...,j_K) \min\left(\sum_{k=1}^{K} j_k g_k, \bar{N}_S\right). \quad (69)$$

The successive minima in these expressions enables to ensure that we never count more slots than $\bar{N}_S$ in the averaging of the numbers of slots occupied by UGS traffic.

Lastly, note that when V_{max}, the limit on the maximum simultaneous UGS calls allowed in the cell is small enough to ensure that no degrading will occur, the expression of $\bar{N}_S^{ugs}(v_1,...,v_R)$ can be drastically simplified as:

$$\bar{N}_S^{ugs}(v_1,...,v_R) = \bar{g}(v_1,...,v_R). \quad (70)$$

6.3.2 Multi-trafic BE

To consider multiple BE traffic profiles, we use the multi-traffic extension presented in Section 5.3 while replacing $\bar{N}_S$ by $\bar{N}_S - \bar{N}_S^{ugs}$ in the different expressions. We then obtain the steady states $\pi_{be}(n_1,...,n_R)$ and the performance parameters in the same way.

7 UGS/ERTPS/BE MODEL

In this section, we first explain how to integrate the traffic of the ertPS service class in our multiclass modeling. Then, we validate the resulting UGS/ertPS/BE model through comparison with simulations.

7.1 Adding ertPS

To integrate the ertPS service class in our UGS/BE model, we follow a similar approach than in Section 6.1.1. Again, no specific scheduling is suggested in the WiMAX standard. However, the QoS needs characterizing each service class lead to consider that ertPS connections preempt BE connections but are preempted by UGS ones [23], [28], [29].

So, our UGS/ertPS/BE model consists in the cascading resolution of the three models, each corresponding to the characterization of the traffic of a given service class. The 3-steps resolution is as follows:

1. We first solve the UGS model to characterize the UGS traffic and compute $\bar{N}_S^{ugs}$, the mean number of slots occupied by the UGS connections.
2. We then solve the ertPS model. Although, this time, we only consider $\bar{N}_S - \bar{N}_S^{ugs}$ available slots in the cell. Similarly to the previous step, we compute the mean number $\bar{N}_S^{ertps}$ of slots occupied by the ertPS connections.
3. Finally, we solve the BE model to obtain the performance parameters of the BE service class using the AGG method as explained in Section 6.1.4. But, we here consider only $\bar{N}_S^{be} = \bar{N}_S - \bar{N}_S^{ugs} - \bar{N}_S^{ertps}$ available slots for the BE connections.

7.2 Validation

To validate our UGS/ertPS/BE model, we now compare its results with those of simulations. To this aim, we use a simulator which implements an ON/OFF traffic generator and a wireless channel for each user. Besides, a centralized scheduler allocates the slots to the active mobiles on a frame by frame basis according to their current MCS and service class. At each frame, the scheduler first serves the UGS connections, then the ertPS connections and lastly the BE connections.

In the simulations, the durations of UGS and ertPS ON and OFF periods are exponentially distributed. In addition, ertPS ON periods are decomposed in talk spurts and silences as recommended in [10].

BE OFF durations are also exponentially distributed. However, contrary to our model, the BE ON sizes are drawn according to a truncated Pareto distribution. Indeed, the truncated Pareto distribution is well known to fit the reality of WEB data traffic. The mean value of the truncated Pareto distribution is given by:

$$\bar{x}_{on} = \frac{\alpha L}{\alpha - 1} \left[1 - (L/H)^{\alpha - 1} \right], \tag{71}$$

where α is the shape parameter, L is the minimum value of Pareto variable and H is the cutoff value for truncated Pareto distribution. For the sake of comparison, we keep the same the mean BE ON size $\bar{x}_{on}$. We set the value of $H = 100\,\bar{x}_{on}$ and consider $\alpha = 1.2$ as suggested in [17]. Finally, we deduce the value of L using relation (71).

The wireless channel parameters are summarized in Table 1. In our analytical model, the channel model is assumed to be memoryless, i.e., MCS are independently drawn from frame to frame for each user. In order to show the robustness of this assumption, we model the channel variations of a simulated mobile by a finite state Markov chain (FSMC) corresponding to a slowly varying Nakagami-m fading channel as proposed in [27]. Each state of the FSMC corresponds to one of the 5 considered MCS (outage included). The state transition matrix C associated to this FSMC is as follows:

$$C = \begin{pmatrix} 0.020 & 0.980 & 0 & 0 & 0 \\ 0.163 & 0.120 & 0.717 & 0 & 0 \\ 0 & 0.277 & 0.620 & 0.103 & 0 \\ 0 & 0 & 0.398 & 0.080 & 0.522 \\ 0 & 0 & 0 & 0.089 & 0.911 \end{pmatrix}, \tag{72}$$

where $C_{i,j}$ is the probability that the MCS of a mobile change from MCS_i to MCS_j. The transitions occur only between adjacent states as we assume in the simulations that the channel is slowly fading. Let us highlight that the steady state probabilities of this FSMC are equal to the p_k probabilities used in the analytical model.

We have repeated comparisons between results from both our analytical models and simulations while considering all sorts of scenarios (e.g., different numbers of traffic profiles in each service classes, different BE scheduling, etc.). Here we present the results of two representative scenarios. In both cases, we observe the behaviors of $\bar{Q}$, $\bar{U}$ and $\bar{X}$ parameters of each service class obtained with our analytical models and compare them with the results of simulations.

7.2.1 First validation scenario

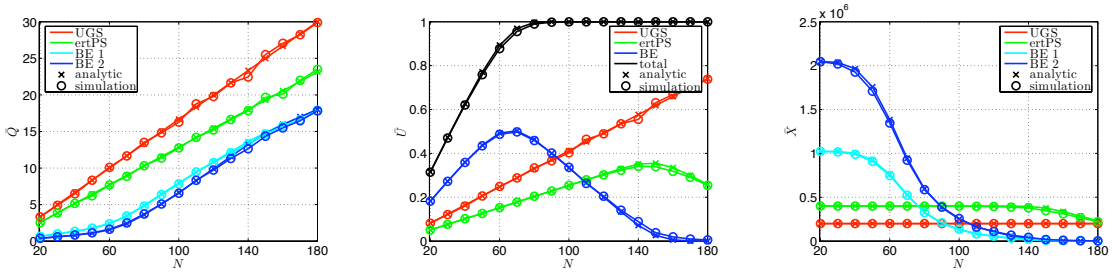
The cell and traffic parameters constituting this first scenario are detailed in Table 3. We assume a total number, N, of mobiles present in the cell ranging from 20 to 180. 50% of these mobiles generates UGS traffic and 30% rtPS traffic. The remaining 20% constitutes the population of BE mobiles in the cell and are equally distributed into two traffic profiles each representing 10% of the total population of mobiles. Finally, note that we consider the throttling policy to allocate slots among the BE connections and that we associate a different $MSTR$ to each BE traffic profile.

Fig. 7(a) presents the average numbers, $\bar{Q}$, of concurrent active connections in the cell. Obviously as the traffic load increases so does these numbers since more and more connections share the limited amount of resource.

The average utilization, $\bar{U}$, of the resource by UGS, ertPS and BE traffics is depicted in Fig. 7(b) and the instantaneous throughput per mobile, $\bar{X}$, is illustrated in Fig. 7(c). At first ($N < 50$), there is always enough resource to satisfy all demands. The parts of the resource occupied by each service class keep on increasing and all connections get their desired throughputs. However, when there are more mobiles in the cell ($N > 50$), the BE mobiles are the first to suffer the lack of resource. As such, their utilization of the frames and their throughputs dive. Finally, observe that when even more mobiles are present in the cell ($N > 150$), even ertPS connections start to see their performances deteriorate. This is explained by the fact that UGS connections are always served first, followed by ertPS connections and then by BE connections.

Parameter		Value	
Number of slots per frame, $\bar{N}_S$		450	
Frame duration, T_F		5 ms	
Limit on UGS calls, V_{max}		50	
Limit on ertPS transfers, W_{max}		50	
BE scheduling		throttling	
Number of mobiles in the cell, N		20 to 180	
UGS	Number of UGS mobiles, N_{ugs}	50% of N	
	Guaranteed bit rate, GBR	200 Kbps	
	Mean ON duration, $\bar{t}_{on}^{ugs}$	60 s	
	Mean OFF duration, $\bar{t}_{off}^{ugs}$	120 s	
ertPS	Number of ertPS mobiles, N_{ertps}	30% of N	
	Guaranteed bit rate, GBR	400 Kbps	
	Mean ON duration, $\bar{t}_{on}^{ertps}$	90 s	
	Mean OFF duration, $\bar{t}_{off}^{ertps}$	120 s	
Traffic profile		**1**	**2**
Number of mobiles per profile		$N_1 = N_2$	
BE	Number of BE mobiles, N_{be}	20% of N	
	Maximum bit rate, $MSTR$	1024 Kbps	2048 Kbps
	Mean ON size, $\bar{x}_{on}^{be}$	3 Mbits	3 Mbits
	Mean OFF duration, $\bar{t}_{off}^{be}$	6 s	6 s

Table 3: First validation scenario.



(a) Mean numbers, $\bar{Q}$, of active UGS, ertPS and BE connections.

(b) Mean utilization, $\bar{U}$, of the resource by UGS, ertPS and BE traffics.

(c) Instantaneous throughput, $\bar{X}$, of a mobile depending on its service class.

Figure 7: First validation scenario: Customary traffic parameters when traffic increases in the three service classes.

It is obvious from the curves depicted in the three figures that the results given by our analytical model match those of simulations. Indeed, the difference between them is less than 4% in most cases and less than 9% in the worst case. This shows the robustness of our model toward the distribution of BE ON sizes and the channel model.

7.2.2 Second validation scenario

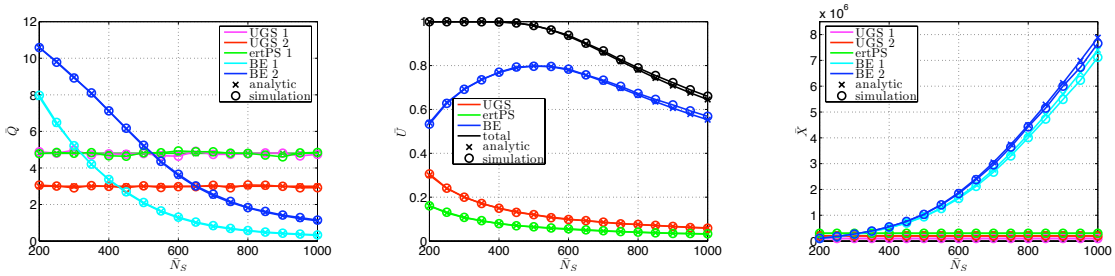
The cell and traffic parameters of the second scenario are presented in Table 4. This time we consider a fixed population of $N = 60$ mobiles in the cell and observe the impact of $\bar{N}_S$ the mean number of available slots on the performance. The numbers of mobiles of each service class are as follows: 24 mobiles generate UGS traffic, 12 ertPS traffic and 24 BE traffic. Also, the UGS and BE connections are equally distributed into two traffic profiles. Lastly, note that the slots are allocated to BE connections

according to the throughput fairness policy.

Parameter		Value	
Number of slots per frame, N_S		200 to 1000	
Frame duration, T_F		5 ms	
Limit on UGS calls, V_{max}		24	
Limit on ertPS transfers, W_{max}		12	
BE scheduling		throughput fairness	
Number of mobiles in the cell, N		60	
Traffic profile		**1**	**2**
UGS	Number of UGS mobiles, N_{ugs}	12	12
	Guaranteed bit rate, GBR	100 Kbps	200 Kbps
	Mean ON duration, $\bar{t}_{on}^{ugs}$	80 s	40 s
	Mean OFF duration, $\bar{t}_{off}^{ugs}$	120 s	120 s
ertPS	Number of ertPS mobiles, N_{ertps}	12	
	Guaranteed bit rate, GBR	300 Kbps	
	Mean ON duration, $\bar{t}_{on}^{ertps}$	80 s	
	Mean OFF duration, $\bar{t}_{off}^{ertps}$	120 s	
BE	Number of BE mobiles, N_{be}	12	12
	Mean ON size, $\bar{x}_{on}^{be}$	1 Mbits	4 Mbits
	Mean OFF duration, $\bar{t}_{off}^{be}$	5 s	5 s

Table 4: Second validation scenario.

Fig. 8(a), 8(b) and 8(c) respectively show the evolution of the $\bar{Q}$, $\bar{U}$ and $\bar{X}$ performance parameters of each traffic profile when $\bar{N}_S$ increases. The mean numbers of active UGS and ertPS mobiles are constant as their ON and OFF durations do not depend on the number of available slots.



(a) Mean numbers, $\bar{Q}$, of active UGS, ertPS and BE connections.

(b) Mean utilization, $\bar{U}$, of the resource by UGS, ertPS and BE traffics.

(c) Instantaneous throughput, $\bar{X}$, of a mobile depending on its service class.

Figure 8: Second validation scenario: Customary traffic parameters when $\bar{N}_S$ the number of available slots increases.

On the contrary, the mean number of active BE connections drop when $\bar{N}_S$ increases. Indeed, more available slots mean faster BE transfers. The UGS and ertPS calls need a given average amount of resource to attain their respective GBR. This constant amount corresponds to a decreasing proportion of the frame as the value of $\bar{N}_S$ rises. When $\bar{N}_S$ is still small ($\bar{N}_S < 500$ slots), the BE mobiles need all the remaining slots of the frames. As such, their mean utilization of the frame grows as more and more slots are left to them by UGS and ertPS calls. On the opposite, when $\bar{N}_S$ is big enough ($\bar{N}_S \geq 500$ slots), the frames are not fully occupied anymore and the proportion of the frame needed for the BE transfers

decreases. Finally, the throughput of BE connections keep on increasing with the number of available slots because they are now scheduled according to the throughput fairness policy.

The results of our analytical model and those of simulations still match each other with the same precision. Here, the difference between them is less than 3% in most cases and less than 7% in the worst case. Once again, the curves show the strong robustness of our model toward the distribution of BE data volumes and the channel model. Indeed, the results of simulations considering a truncated Pareto distribution of BE ON sizes and a slowly fading channel match very closely those provided by our analytical model.

Besides, note that the results presented in this section are representative of the results obtained for the numerous scenarios we have considered. Indeed, each time, the simulations results validated our analytical models with similar accuracy.

At last, let us highlight that the simulations required very long computation durations (nearly a day) whereas the results of the analytic models where obtained instantaneously.

8 DIMENSIONING

In this section, we provide dimensioning examples to demonstrate possible applications of our models. We study the case of an operator wishing to dimension a WiMAX cell. This operator considers subscribers distributed in two profiles, business and domestic, corresponding to their utilization of the cell as detailed in Table 5.

Parameter		**Value**	
Number of slots per frame, N_S		450	
Frame duration, T_F		5 ms	
Limit on UGS calls, V_{max}		60	
Limit on ertPS transfers, W_{max}		30	
BE scheduling		opportunistic	
Subscriber profile		**Business**	**Domestic**
Number of subscribers in the cell		10 to 150	
Proportion of subscribers		50%	50%
UGS	Guaranteed bit rate, GBR	128 Kbps	64 Kbps
	Mean ON duration, $\bar{t}_{on}^{ugs}$	120 s	60 s
	Mean OFF duration, $\bar{t}_{off}^{ugs}$	180 s	180 s
ertPS	Guaranteed bit rate, GBR	256 Kbps	no domestic ertPS traffic
	Mean ON duration, $\bar{t}_{on}^{ertps}$	60 s	
	Mean OFF duration, $\bar{t}_{off}^{ertps}$	180 s	
BE	Mean ON size, $\bar{x}_{on}^{be}$	varying	
	Mean OFF duration, $\bar{t}_{off}^{be}$	varying	

Table 5: Cell and traffic parameters.

The channel parameters are provided in Table 1.

Business subscribers are assumed to generate important UGS, ertPS and BE traffics. As such, a business subscriber is represented in our model by three mobiles, one for each service class, and the traffic he generates is the sum of the traffic generated by the three mobiles. On the opposite, domestic subscribers generate a lower amount of UGS, an equivalent amount of BE traffics and no ertPS traffic at all. So, a domestic subscriber is only represented in our model by two mobiles: one mobile for UGS and one for BE. In the following examples, we consider the same proportion of business and domestic

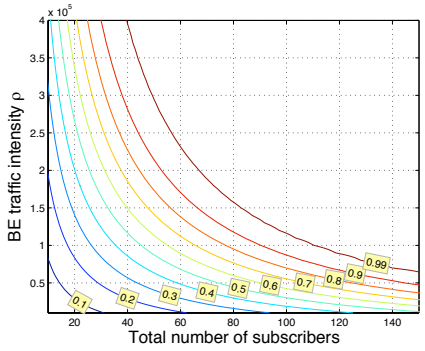
subscribers present in the cell.

Let us highlight that results can be obtained for any other possible configuration (i.e., any mono or multi-profile traffic scenario) by using the according models.

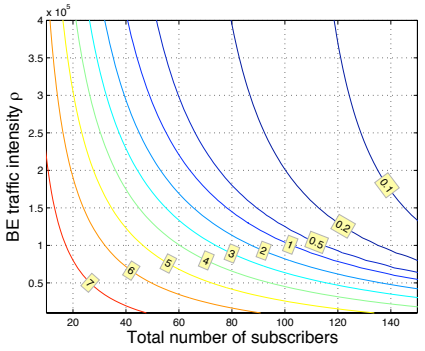
8.1 Performance graphs

Here, we study contour graphs in which a performance parameter is drawn as a function of the parameters to dimension, e.g., the total number of subscribers in the cell, the number of available slots, $\bar{N}_S$, traffic intensities, etc.

The mean radio resource utilization of the WiMAX cell $\bar{U}$, and the average throughput of a BE connection $\bar{X}_{be}$ are illustrated in Fig. 9(a) and 9(b).



(a) Mean utilization, $\bar{U}$, of the resource.



(b) Instantaneous throughput, $\bar{X}_{be}$, of a BE connection (in Mbps).

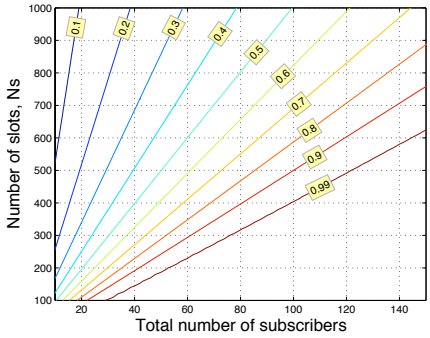
Figure 9: $\bar{U}$ and $\bar{X}_{be}$ traffic parameters when the population of subscribers in the cell and the BE traffic intensity increase.

These parameters are presented as functions of the total number of subscribers in the cell and ρ_{be}, the intensity of the traffic generated by the BE mobiles. ($\rho_{be} = \frac{\bar{x}_{on}^{be}}{\bar{t}_{off}^{be}}$ as described in Section 5.2.) To determine $\bar{U}$ or $\bar{X}_{be}$ for a given number of subscribers in the cell and a given BE traffic intensity, we first locate the point at these coordinates in the corresponding graph. This point is located between two contour lines. The line with the higher value gives an upper bound on the parameter value and the line with the smaller value gives a lower bound. For example, 40 subscribers in the cell and a BE traffic intensity of 200 Kbps (e.g., $\bar{x}_{on}^{be} = 2$ MB and $\bar{t}_{off}^{be} = 10$ s) lead to $0.7 < \bar{U} < 0.8$ and 3 Mbps $< \bar{X}_{be} < 4$ Mbps.

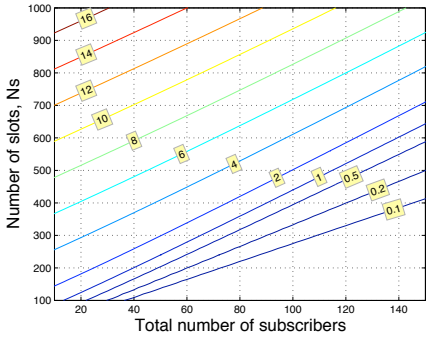
As depicted on these graphs, the mean utilization $\bar{U}$ of the frame rises and the instantaneous throughput of the BE connections $\bar{X}_{be}$ drops when the number of subscribers and/or the BE traffic intensity increase.

Similarly, Fig. 10(a) and 10(b) show the same performance parameters but this time as functions of both the number of available slots and the total number of subscribers in the cell. ρ_{be} is set to 200 Kbps. We can observe that low frame utilizations and important BE throughputs are achieved when high amounts of slots are available and only small numbers of subscribers share the resource.

Each graph is the result of several thousands of input parameter sets. Obviously, any simulation tool or even any multidimensional Markov chain requiring numerical resolution, would have precluded the drawing of such graphs.



(a) Mean utilization, $\bar{U}$, of the resource.

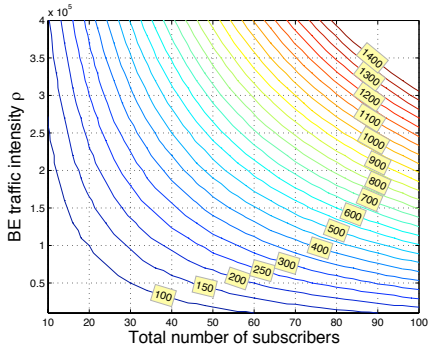


(b) Instantaneous throughput, $\bar{X}_{be}$, of a BE connection (in Mbps).

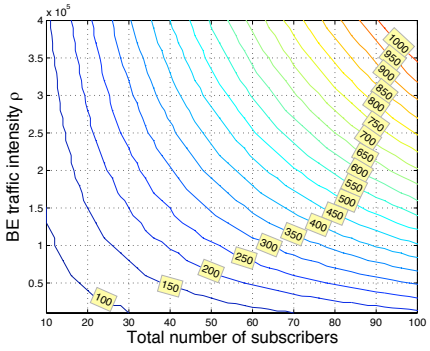
Figure 10: $\bar{U}$ and $\bar{X}_{be}$ traffic parameters when the population of subscribers in the cell and the number N_S of available downlink slots per frame increase.

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 presented in Fig. 11(a) and 11(b). To draw these two graphs, we computed the considered performance parameter ($\bar{U}$ or $\bar{X}_{be}$) for each possible number of subscribers and ρ_{be} pair while increasing the number N_S of available slots until the chosen QoS criterion could not be guaranteed anymore. Note that this straightforward method is only rendered possible due to the instantaneous resolution aspect of our model.



(a) Maximum number of slots, N_S^{max} to guarantee $\bar{U} \geq 0.85$.



(b) Minimum number of slots, N_S^{min} to guarantee $\bar{X}_{be} \geq 1024$ Kbps.

Figure 11: Dimensioning of the number N_S of available downlink slots per frame.

In Fig. 11(a) we find the maximum number $\bar{N}_S^{max}$ of available slots guaranteeing an average radio utilization over 85%. This kind of criterion allows operators to avoid uneconomical over-provisioning of the network resources in regard to the traffic load of their customers. To obtain the optimal value of $\bar{N}_S^{max}$ associated with a number of subscribers and a given value of ρ_{be}, we look for the point at the corresponding coordinates in the graph. This point is located between two contour lines, and the one with the lower value gives the value of $\bar{N}_S^{max}$.

The QoS criterion chosen as a second example is the throughput per BE connection. We observed BE traffic throughputs because BE connections are served in last. As such, guaranteeing a minimum throughput for BE connections leads to ensure that UGS and rtPS connections obtain their reserved bit rate. We decided on 1024 Kbps as the arbitrary value of the minimum throughput of a BE connection. Now, we want to find the minimum number $\bar{N}_S^{min}$ of available slots guaranteeing this minimum throughput

threshold. In Fig. 11(b), a given point is located between two contour lines. The line with the higher value gives $\bar{N}_S^{min}$.

The graphs of Fig. 11(a) and 11(b) can be jointly used to satisfy multiple QoS criteria. For example, if we consider a WiMAX cell with 50 subscribers and a BE traffic intensity of 200 Kbps, Fig. 11(a) gives $\bar{N}_S^{max} = 500$ slots, and Fig. 11(b) gives $\bar{N}_S^{min} = 350$ slots. The combination of these two results recommends to have a number of slots $\bar{N}_S \in [350; 500]$ to guarantee both a reasonable resource utilization and an acceptable minimum throughput to the subscribers. This roughly corresponds to a system bandwidth between 8 and 10.5 MHz if we consider a downlink/uplink ratio of 2/3, a Fast Fourrier Transform size equal to 2048, a PUSC subcarrier permutation and an average protocol overhead length of 4 symbols.

9 CONCLUSION

As deployment of 4G networks is underway, need arises for operators and manufacturers to develop dimensioning tools. In this paper, we have designed useful analytical models for the three service classes defined in the WiMAX standard: UGS, ertPS and BE, respectively corresponding to VoIP, VoIP with silence suppression and best-effort traffic. Also, we have explained how these models could be easily adapted for LTE systems.

As a first step, we presented models considering cells carrying either UGS, ertPS or BE traffic. For each model, we proposed multi-traffic extensions in order to account for users generating traffics of different intensities. Besides, we provided closed-form expressions of all the required performance parameters.

Then, we have detailed three methods to combine the UGS and BE models. The results obtained with each method proved to be very close. We deduced that averaging the numbers of slots occupied by the traffics of the different service classes was sufficient to combine the models. The resulting UGS/ertPS/BE model, based on the cascading resolution of the UGS, ertPS and BE models, is able to instantaneously evaluate the performance parameters of each service class. This multi-class model can even take into account users of the same service class generating traffics of different profiles with a minimal increase of its resolution complexity. Therefore it renders possible efficient and advanced dimensioning studies as shown in this paper.

We used extensive simulations to validate our analytical approach. To show the robustness of our traffic and channel assumptions, we compared our model's results to simulations considering a slowly fading channel and a truncated Pareto distribution of BE ON sizes. The simulation results showed the accuracy and robustness of our analytical modeling: for all considered scenarios, maximum relative errors never exceeded 9%.

Lastly, let us highlight that we focused in this paper on VoIP and best-effort traffic. However, another type of traffic is expected to be supported by 4G networks: streaming. The streaming connections need guarantees on both their minimum bit rate and the maximum latency of their packets. This last constraint renders the modeling of streaming traffic very difficult. In our next work, we will tackle the challenging task of developing a model dedicated to streaming in OFDMA-based cellular network. Ultimately, our aim will be to integrate this model in our cascade and thus be able to instantaneously obtain results for all types of traffic.

REFERENCES

[1] 3GPP LTE: Homepage with specifications - http://www.3gpp.org/LTE.

[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] 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.

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

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

[8] S. Borst and N. Hedge. Integration of Streaming and Elastic Traffic in Wireless Networks. In *IEEE Infocom*, 2007.

[9] P. T. Brady. A technique for investigating ON/OFF patterns for speech. In *Bell Labs Syst. Tech. J., vol. 44, no. 1*, pages 1–22, 1965.

[10] P. T. Brady. A model for generating ON/OFF speech patterns in two-way conversations. In *Bell Labs Syst. Tech. J., vol. 48, no. 9*, pages 2445–2472, 1969.

[11] K.-C. Chen and J. R. B. de Marca. *Mobile WiMAX*. Wiley, February 2008.

[12] F. Delcoigne, A. ProutiŕRre, and G. RŐgniŐ. Modelling integration of streaming and data traffic. In *Performance Evaluation*, 2004.

[13] M. Dirani, C. Tarhini, and T. Chahed. Cross-layer modeling of capacity in wireless networks Application to UMTS-HSDPA, IEEE802.11 WLAN and IEEE802.16 WiMAX. In *Computer Communications*, 2007.

[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] S. Doirieux, B. Baynat, M. Maqbool, and M. Coupechoux. An Efficient Analytical Model for the Dimensioning of WiMAX Networks Supporting Multi-profile Best Effort Traffic. *Computer Communications Journal, to appear*, 2010.

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

[17] 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.

[18] 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.

[19] C. Hoymann. Analysis and performance evaluation of the OFDM-based metropolitan area network IEEE 802.16. In *Computer Networks*, 2005.

[20] M. Ibrahim, K. Khawam, A. E. Samhat, and S. Tohme. Analytical framework for dimensioning hierarchical WiMax-WiFi networks. In *Computer Networks*, 2009.

[21] V. B. Iversen. *Teletraffic engineering and network planning*. Technical University of Denmark, 2006.

[22] A. Jensen. Truncated multidimensional distributions. *The Life and Works of A.K. Erlang*, pages 58–70, 1948.

[23] C.-H. Jiang and T.-C. Tsai. Token bucket based CAC and packet scheduling for IEEE 802.16 broadband wireless access networks. In *Proc. of the 3rd IEEE Consumer Communications and Networking Conference*, January 2006.

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

[25] J. Lu and M. Ma. A cross-layer elastic CAC and holistic opportunistic scheduling for QoS support in WiMAX. In *Computer Networks*, 2010.

[26] M. Maqbool, M. Coupechoux, P. Godlewski, S. Doirieux, B. Baynat, and V. Capdevielle. Dimensioning Methodology for OFDMA Networks. In *Proc. of the Wireless World Research Forum (WWRF22)*, 2009.

[27] 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.

[28] A. Sayenko, O. Alanen, J. Karhula, and T. Hamalainen. Ensuring the QoS requirements in 802.16 scheduling. In *Proc. of the 9th ACM International Symposium on Modeling Analysis and Simulation of Wireless and Mobile Systems*, 2006.

[29] V. Singh and V. Sharma. Efficient and fair scheduling of uplink and downlink in IEEE 802.16 OFDMA networks. In *Proc. of the IEEE Wireless Communications and Networking Conference*, 2006.

[30] C. Tarhini and T. Chahed. On Mobility of Voice-Like and Data Traffic in IEEE802.16e. In *Global Telecommunications*, 2008.

[31] Z. Yumei and S. Yu. Scheduling Algorithm with Quality of Service Support in IEEE 802.16 Networks. In *Computer Networks*, 2009.