Analytical Modelling and Message Delay Performance Evaluation of the IEEE 802.16 MAC Protocol

Luís F. M. de Moraes and Danielle Lopes F. G. Vieira
Laboratório de Redes de Alta Velocidade - Universidade Federal do Rio de Janeiro (UFRJ)
Email: moraes,danielle@ravel.ufrj.br

***Abstract*— The IEEE 802.16 is a well known set of standards, which have been developed for global deployment of Metropolitan Area Networks (MANs) in order to provide broadband wireless accesses supporting the integrated transmission of multimedia applications with different Quality of Service (QoS) requirements. In order to achieve the QoS requirements of multimedia applications, the IEEE 802.16 standards offers different scheduling schemes. This paper is concerned with the analytical modelling of the MAC protocol and the analysis of average message delays for real and non-real time traffics in broadband access wireless networks operating under the IEEE 802.16 standard. The message delays results obtained are presented for two scenarios, as a function of the load generated for different types of traffic. In addition, the impact on the network performance caused by some of the parameters utilized as part of the random access mechanism is also investigated. The results obtained through the use of the analytical model proposed in this paper are compared with those obtained with the aid of a simulation tool. The agreement of comparisons involving analytical and simulation results in the examples studied show that the model proposed in this paper could be very useful when utilized to represent and study further behavior characteristics of the end-to-end message delays involved in IEEE 802.16 standard MAC protocol under consideration.**

I. Introduction

The increasing demand for high-speed Internet access and multimedia services provided fast wireless access development for wireless metropolitan area networks (WMAN). The IEEE 802.16 standard [1] came to provide high-speed wireless access and high performance, with service differentiation for different types of traffic with different QoS requirements.

To provide this QoS, the IEEE 802.16 standard Media Access Control (MAC) protocol offers different options for sharing the wireless medium, for different types of traffic. However, it is known that one of the main challenges in network communications is how to find an economical and efficient form of sharing the transmission medium [2]. The IEEE 802.16 standard offers two solutions for multiple access in the *uplink* [1]. The first is to allocate resources for each station, whether or not they have data waiting to be sent. The other possibility is just to allocate resources for the station when it is ready to send data and explicitly request bandwidth for it. The protocol in IEEE 802.16 standard MAC layer uses these two solutions to allocate resources for different traffic types.

Authors in alphabetical order of last names.

The performance analysis of the IEEE 802.16 standard was studied extensively in the literature. However, those papers presented performance evaluations through results obtained by simulation. There are papers that focus on analytical modelling, however these also propose changes to the standard MAC protocol. Besides, to the best of our knowledge, no analytical model for the total delay of messages was presented for this IEEE standard until now.

This article proposes an analytical model for message delay that takes into account the characteristics of the IEEE 802.16 standard, for users with different service requirements and for types of traffic demanding distinct QoS. The model proposed here allows performance evaluation of the IEEE 802.16 standard under the metric of total message delay. These results, obtained through the proposed analytical model, were compared with results obtained through simulations.

To the best of our knowledge, this type of analysis has not yet been made for IEEE 802.16 standard. More specifically, previous performance evaluation work on IEEE 802.16 focuses on specific aspects. In [3], the authors described the QoS framework of 802.16 and discussed simulation results in specific application scenarios. In [4], authors presented a simulation study of the IEEE 802.16 MAC protocol and they evaluated the behavior of the system over different scenarios of traffic and varying the values of a set of system parameters. Aura Ganz *et al.* [5] proposed a packet scheduler for the IEEE 802.16 uplink, based on a hierarchical structure of queues, and they developed a simulation model to evaluate the behavior of the proposed scheduler. A new QoS architecture for IEEE 802.16, where the scheduling is based on the packet life time of each flow type, has been proposed in [6]. An stochastic analysis to find the best reservation period size, with the intention of optimizing medium access has been proposed in [7]. The performance with Time-Division Duplexing (TDD) mode has been analyzed in [8]. In [9], authors studied through simulation the behavior of different random access mechanisms presented in the IEEE 802.16 standard. In [10], the behavior of the IEEE 802.16 MAC protocol's upstream frame was investigated through simulations, in terms of the flow, average delay and probability of message collisions. In [11], authors proposed a MAC protocol based on polling and presented an analytical model to evaluate its behavior in terms of message delay.

The remainder of this paper is organized as follows. Section

II describes briefly the IEEE 802.16 standard. Sections III and IV introduces the proposed analytical model and the obtained results respectively. Finally, conclusions and considerations for future work are discussed in Section V.

II. IEEE 802.16 STANDARD

The IEEE 802.16 standard specifies a wireless interface for WMANs and it has been developed with the purpose of standardizing broadband wireless access technology, defining the wireless interface and the MAC protocol for WMANs.

The network architecture that uses the IEEE 802.16 standard has two main elements: Base Station - BS, that coordinates all the communication, and Subscriber Station (SS), that are located at different distances from the BS, in a Point-Multi-point (PMP) topology. Transmissions happen in two different channels: a downlink channel (DL), with the addressed data flow from the BS to the SSs and transmitted by broadcast, and another of uplink (UL), with the data flow addressed from SSs to BS, where the SSs share the media.

The duplexing between these architecture elements can happen in two different ways: Frequency-Division Duplexing (FDD) and Time-Division Duplexing (TDD). In the TDD, during the DL, data packets are transmitted by diffusion from BS to all SSs, that just keep the packets destined to them. During the uplink, through the message UL-MAP at the beginning of each frame, BS broadcasts the number of segments that will be attributed for each SS of the sub-frame. In order to send grant requests to the BS, SSs use a random access and piggybacking [1] in the uplink frame [1]. For collision resolution during this interval, the standard defines a backoff algorithm.

The MAC layer also provides mechanisms to ensure QoS for different types of traffic. The main mechanism for QoS provision consists of associating the packets transmitted by the MAC layer to a given service flow. Each service flow should define its group of QoS parameters, such as maximum delay, minimum bandwidth and the type of scheduling service. The standard specifies four scheduling services, where each flow is associated to one of those services and the BS scheduler allocates bandwidth for SSs following the group of rules defined by each service [1].

The first of the scheduling services is *Unsolicited Grant Service* (UGS) is designed to support real-time applications, with strict delay requirements. Grants occur on a periodic basis. The base period and the grant size are specified during the connection setup phase. *Real-Time Polling Service* (rtPS) is designed to support real-time applications with less stringent delay requirements, which generate variable-size data packets at periodic intervals. The BS periodically sends unicast polls to rtPS connections. The base period can be specified during the connection setup. *Non-real-time Polling Service* (nrtPS) and *Best Effort* (BE) are designed for applications that do not have specific delay requirements. The main difference between them is that nrtPS connections are reserved a minimum amount of bandwidth. Both nrtPS and BE uplink connections typically use contention-based bandwidth requests. Such requests are sent in response to broadcast/multicast polls, which are advertised by the BS in the UL-MAP.

[1]Requests sent by SSs in the end of the data frame, transmitted during the uplink.

III. ANALYTICAL MODEL

To evaluate the performance of the IEEE 802.16 MAC protocol, this section introduces an analytical model that incorporates the characteristics of this protocol. As seen in the previous section, IEEE 802.16 networks are regulated by a request-grants mechanism in the process of bandwidth request and for resource grants that are pre-negotiated in the case of higher priority traffic. Therefore, the analysis is divided into two distinct parts: bandwidth request phase (collision resolution) and data transmission phase.

A. Delay in the Bandwidth Request Phase

This section commences by constructing a model for the backoff algorithm and then discusses the behavior of a single station within the proposed Markov Chain model. The IEEE 802.16 method for contention resolution is based on a backoff algorithm, with the initial and maximum values of the backoff window controlled by BS. For details on the algorithm see [1]. The technique used to obtain the average message delay for the bandwidth request is similar to the method employed in [12].

Consider a fixed number N of contending stations. In saturation conditions, each station has immediately a packet available for transmission, after the completion of each successful transmission. Let $B(t)$ be the stochastic process representing the backoff time counter for a given station. The backoff counter, k, is defined as the number of contention transmission opportunities for which the station must wait prior to commencing transmission. Let m, maximum backoff stage, be the value such that $W_{max} = 2^m W_{min}$, where W_{min} is the initial value of the backoff window, and let us adopt the notation $W_i = 2^i W_{min}$, where $i \in (0, m)$ is called backoff stage. Let $S(t)$ be the stochastic process representing the backoff stage $(0, ..., m+R)$ of the station at time t, where R represents the retry backoff stage.

Figure 1 represents a two-dimensional process $\{S(t), B(t)\}$ in the form of a discrete-time Markov Chain. In this figure, the backoff period unit is expressed by W and is of the same size as the slot time. p represents the conditional collision probability, which is an independent event probability with a constant value.

It is known that the model is irreducible, aperiodic and recurrent non-null. Hence, a stationary distribution of the current model exists. From the seven single-step transition probabilities defined for the chain, the stationary distribution, $b_{i,k}$ ($b_{i,k} = \lim_{t\to\infty} P_r\{S(t) = i, B(t) = k\}$ if $i \in (0, m)$ and $k \in (0, W_i - 1)$ or $b_{i,k} = \lim_{t\to\infty} P_r\{S(t) = i, B(t) = k\}$ if $i \in (m+1, m+R)$ and $k \in (0, W_m - 1)$), can be expressed as:

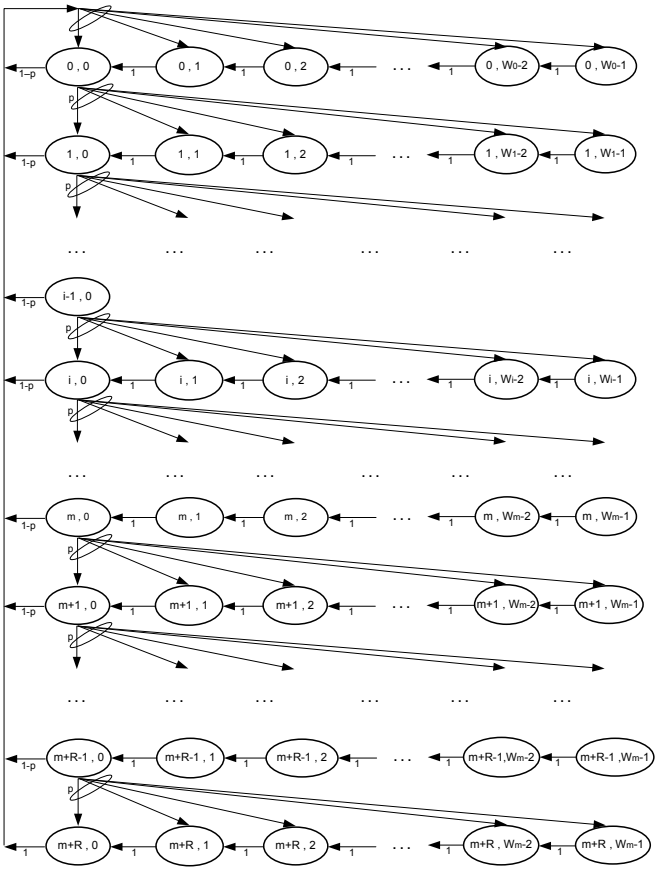


Fig. 1. Markov chain Model for backoff algorithm

$$b_{i,k} = \begin{cases} \frac{W_i - k}{W_i} \begin{cases} (1-p)\sum_{l=0}^{m+R-1} b_{l,0} + b_{m+R,0} & i=0 \quad k \in (0, W_i - 1) \\ p \cdot b_{i-1,0} & 0 < i \leq m k \in (0, W_i - 1) \end{cases} \\ \frac{W_m - k}{W_m} \cdot p \cdot b_{i-1,0} \qquad m < i \leq m+R \quad k \in (0, W_m - 1) \end{cases} \tag{1}$$

where:

$$b_{i-1,0} \cdot p = b_{i,0} \rightarrow b_{i,0} = p^i \cdot b_{0,0} \tag{2}$$

Using Equation 2, Equation 1 can be rewriten as:

$$\begin{cases} b_{i,k} = \frac{W_i - k}{W_i} b_{i,0} & i \in (0, m) \quad k \in (0, W_i - 1) \\ b_{i,k} = \frac{W_m - k}{W_m} b_{i,0} & i \in (m+1, m+R) \quad k \in (0, W_m - 1) \end{cases} \tag{3}$$

Thus, by Equations 2 and 3, all the values $b_{i,k}$ are expressed as functions of the value $b_{0,0}$ and of the conditional collision probability p.

Due to chain regularities, a closed-form solution for this Markov chain can be readily obtained. $b_{0,0}$ is determined by imposing the probability conservation law, i.e.

$$\begin{aligned} 1 &= \sum_{i=0}^{m} \sum_{k=0}^{W_i - 1} b_{i,k} + \sum_{i=m+1}^{m+R} \sum_{k=0}^{W_m - 1} b_{i,k} \\ &= \sum_{i=0}^{m} b_{i,0} \frac{W_i + 1}{2} + \sum_{i=m+1}^{m+R} b_{i,0} \frac{W_m + 1}{2} \qquad (4) \\ &= \frac{b_{0,0}}{\frac{W_{min}[(1-p)(1-(2p)^{m+1}) + (1-2p)2^m p^{m+1}(1-p^R)] + (1-2p)(1-p^{m+R+1})}{(1-2p)(1-p)}} \qquad (5) \end{aligned}$$

where

$$b_{0,0} = \frac{2(1-2p)(1-p)}{W_{min}[(1-p)(1-(2p)^{m+1}) + (1-2p)2^m p^{m+1}(1-p^R)] + (1-2p)(1-p^{m+R+1})} \tag{6}$$

Given the values of R, W_{min} and p, the steady state probability of the model can be calculated from Equations 6 to 6.

Let τ be the steady state probability of a station sending a transmission during any slot time. In the network, a station only transmits when its backoff counter is equal to zero (i.e. the station transmits at any i of $b_{i,0}$).

$$\tau = \frac{2(1-2p)(1-p^{m+R+1})}{W_{min}[(1-p)(1-(2p)^{m+1}) + (1-2p) \cdot 2^m \cdot p^{m+1}(1-p^R)] + (1-2p)(1-p^{m+R+1})} \tag{7}$$

A collision occurs when two or more stations transmit during the same slot time. So, the collision probability, p, of a transmitted request message is given by:

$$p = 1 - (1-\tau)^{N-1} \tag{8}$$

Eqs. 7 and 8 represent a nonlinear system with two unknown parameters, i.e. τ and p. Solving this equation for τ, yields the probability p and enables the subsequent derivation of the stationary distribution by substituting $b_{0,0}$ and p in Eq. 1.

It is still necessary to define some parameters for the reservation mechanism's delay calculation. Let P_{tr} be a probability that there is at least one transmission in the considered time slot. Since N stations contend for the channel, and each transmits with probability τ,

$$P_{tr} = 1 - (1-\tau)^N \tag{9}$$

The probability P_s that a transmission occurring on the channel is successful is given by the probability that exactly one station transmits on the channel, conditioned on the fact that at least one station transmits, i.e.:

$$P_s = \frac{\binom{N}{1} \tau (1-\tau)^{N-1}}{P_{tr}} = \frac{N\tau(1-\tau)^{N-1}}{1-(1-\tau)^N} \tag{10}$$

The average delay of the bandwidth request messages $E[D_r]$, is defined as the elapsed time between its generation and its successful reception. Collisions may occur during the transmission process, therefore:

$$E[D_r] = E[N_c](E[\delta] + T_c) + (E[\delta] + T_s) \tag{11}$$

where $E[N_c]$ is the expected value of number of collisions experienced by a request message before successful reception by BS, $E[\delta]$ is the average time delay of the backoff counter specified by a station before accessing the channel under busy conditions, T_c is the time duration of the collision and finally T_s is the elapsed time for a successful transmission.

From the behavior of a transmission (i.e. it collides continually before successful receipt) and the definition of mean value, it is know that the random variable N_c conforms to a geometric distribution with a parameter P_s. The mean value of N_c is given by:

$$E[N_c] = \sum_{i=1}^{\infty} i(1-P_s)^i P_s = \frac{1}{P_s} - 1 \tag{12}$$

When the station's counter is at state $b_{i,k}$, a time interval of k slots is required for the counter to reach state $b_{i,0}$. This interval is denoted by the random variable β, whose mean value is given by:

$$\begin{aligned} E[\beta] = & \sum_{i=0}^{m} \sum_{k=1}^{W_i} k b_{i,k} + \sum_{i=m+1}^{m+R} \sum_{k=1}^{W_{m-1}} k b_{i,k} \quad \cdot \\ = & \frac{b_{0,0}}{6} \cdot \\ & \left\{ \frac{W_{min}^2[(1-p)(1-(4p)^{m+1})+4^m(1-4p)p^{m+1}(1-p^R)]-(1-4p)(1-p^{m+1})+p^{m+1}(1-p^R)}{(1-4p)(1-p)} \right\} \end{aligned}$$

The time that the station's counter remains frozen is denoted by Φ. When the counter freezes, it remains inactive for the duration of one period reserved for data transmission. Then, it can be shown that $E[\Phi] = (E[\beta]/\mu)(E[T_q]-\mu)$, where: $E[T_q]$ is defined as the average time duration of a 802.16 frame and μ is considered a constant parameter and it is defined as the size of the reservation period. Therefore, $E[\delta] = E[\beta] + E[\Phi]$. This permits the mean data message delay time $E[D_r]$ to be calculate.

B. Delay in the Data Allocation Phase

This section proposes a model for data allocation with two priority classes of traffic that supports real-time service, transmitted by UGS, and non-real-time service, transmitted through reservation mechanism or piggybacking. The real-time service has pre-emptive head-of-line (HOL) priority over non-real-time traffic.

As seen in the previous section the bandwidth of a UGS is pre-determined and the BS knows its bandwidth size and service time. Thus, it can be assumed that there are real-time virtual requests whose requested size is equal to the UGS grants and its arrival time is equal to the UGS nominal grant time. The BS has a virtual buffer to accommodate real-time virtual requests (RTVR) and non-real-time requests (NRTR). The arrival of RTVR and NRTR follows an independent Poisson process with arrival rates λ_1 and λ_2, respectively. The service time of RTVR and NRTR is assumed to be independent and identically distributed with a general distribution. Let ν_1 and ν_2 be the mean service times of RTVR and NRTR, respectively. The service disciplines for both traffic are First-Come-First-Served (FCFS). It is considered that μ reservation slots are allocated by the BS in each UL-MAP.

The real-time traffic delay is defined as the time between the actual grant time and the nominal grant time. After arriving at the BS, a NRTR will wait in the buffer until the BS controller allocates bandwidth for it in following UL-MAP. The non-real-time traffic delay is defined as the time between when a NRTR arrival at the BS and when the last bit of the requested data packet arrives at the BS.

1) A model for Analysis: The time scheduling defined by the UL-MAP is demonstrated in Figure 2(a). Assume that the time schedule defined by the last UL-MAP begins at time t_1 and ends at time t_2. The BS receives NRTRs from the stations from time t_1 until time t_2. At time t_2, the BS allocates slots for RTVR, NRTR and contention slots to form a UL-MAP and grant the transmission opportunities.

A model which is a variant of the prioritized buffered leaky Bucket model [13] is created as shown in Figure 2(b). There is a token pool in the model. The size of the token pool is 4096, each token stands for a slot and accordingly a corresponding time. Initially, the token pool is full of 4096 tokens. There is a virtual buffer to accommodate RTVR, a buffer to accommodate the arriving NRTR and another virtual buffer to accommodate μ contention slots. The number of tokens in the token pool is decremented by one for every slot allocated. According to the mechanism of bandwidth request, NRTR arriving from t_1 to t_2 will be served in the next UL-MAP, so the starting time of the NRTR buffer is t_2. RTVR arriving during the time defined by next UL-MAP will be served in the next UL-MAP, so the starting time of the RTVR buffer is t_3.

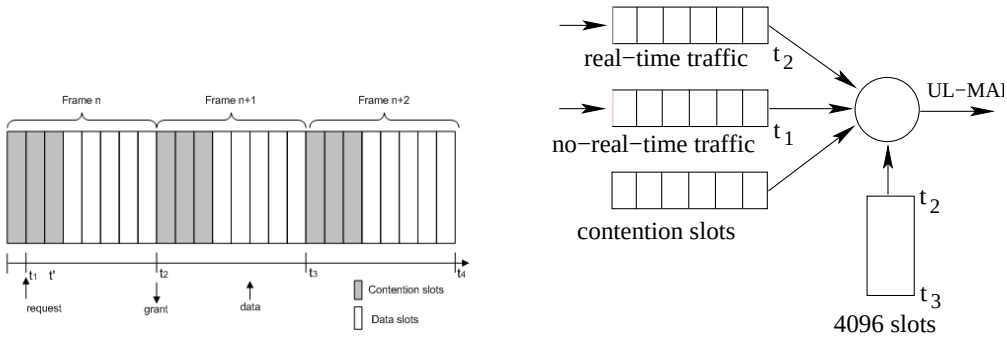
2) Average Delay of Real-Time Traffic: According to the assumptions and the analytical model presented, the slot allocation for NRTR and contention slots will be interrupted by the arrival of real-time virtual requests. Since RTVR have preemptive priority, the process $\{x_t, t \in [0,\infty)\}$, where x_t, is the number of RTVR present in the system at time t, is the queuing process M/G/1. All probabilities relating to x_t can be immediately obtained. The average delay of real-time traffic can be easily found [14], and can be observed in Equation 13:

$$E[D_1] = \frac{\lambda_1 E[V^2]}{2(1-\lambda_1 E[V])} \tag{13}$$

where λ_1 is the arrival rate of RTVR and V is the service time of RTVR. Therefore $E[D_1]$ is the average delay of real-time traffic.

3) Average delay of Non-Real-Time Data Traffic: Assume that the number of NRTR in the non-real-time buffer is $l-1$ and the lth NRTR arrives at the BS at time $t' \in [t_1, t_2]$. The NRTR are served in order, during which the service will be interrupted and RTVR will be served if they arrive. Assume that a number g of RTVRs arrive before the l-th NRTR is served. According to the service discipline and time scheduling scheme defined in Section III-B.1, the delay of the i-th NRTR will be:

$$D_2 = t_2 - t' + \sum_{i=1}^{l} S_i + \sum_{j=1}^{g} V_j \tag{14}$$



(a) Time relation of MAP.

(b) Analytical Model

Fig. 2. Model in the Data Allocation Phase.

where S_i is the random variable that represents the service time for the i-th NRTR, V_j is the random variable that represents the service time for the j-th RTVR, t_2 is the start time of the next UL-MAP and t' is the arrival time of the l-th NRTR. Then the average delay of non-real-time traffic is:

$$E[D_2] = E[T_{MAP}] - E[t^{'}] + E[l]E[S] + E[g]E[V] \quad (15)$$

where $E[T_{MAP}]$ is the average time defined by the UL-MAP, $E[S]$, $E[V]$ are the average service times of non-real-time and real-time traffic respectively. $E[t^{'}]$ is the average arrival time of a NRTR within $[t_1, t_2]$. $E[l]$ is the average number of NRTR arrivals from t_1 to $E[t^{'}]$, and $E[g]$ is the average number of RTVR arrivals which arrive before the i-th NRTR is served.

Under stationary conditions, the time defined by the next UL-MAP consists of the service time of RTVR arriving during the next UL-MAP time, the service time of NRTR arriving from t_1 to t_2 and the time occupied by μ contention slots. That is,

$$\mu + \sum_{i=1}^{G} V_i + \sum_{j=1}^{L} S_j = T_{MAP} \quad (16)$$

where T_{MAP} is the time defined by the next UL-MAP, G is the number of NRTR arriving during the next UL-MAP time and L is the number of RTVR arriving from t_1 to t_2. Here time is measured in units of time slots. Taking the expectation of Equation 16 and substituting $E[G] = \lambda_1 E[T_{MAP}]$, $E[L] = \lambda_2 E[T_{MAP}]$, it is possible to show that $E[T_{MAP}]$ is given by:

$$E[T_{MAP}] = \frac{\mu}{1 - (\lambda_1 E[V] + \lambda_2 E[S])} \quad (17)$$

NRTR arrivals behave as a Poisson process and t' is uniformly distributed [14], hence:

$$E[g] = \frac{\lambda_1 \lambda_2 E[T_{MAP}] E[S]}{2(1 - \lambda_1 E[V])} \quad (18)$$

$$E[l] = \lambda_2 E[T_{MAP}]/2 \quad (19)$$

$$E[t^{'}] = E[T_{MAP}]/2 \quad (20)$$

Using the expected value of Equation 16 and the values defined in Equations 18, 19 and 20 into Equation 15, it is possible to obtain the average delay of non-real-time traffic $E[D_2]$.

In this section we present an analytical model and the performance analysis of the IEEE 802.16 standard. This model allows the calculation of important performance metrics, such as queue size and request message delay as well as data message delay. Numeric results can be obtained easily from the equations that model the system.

IV. Obtained Results

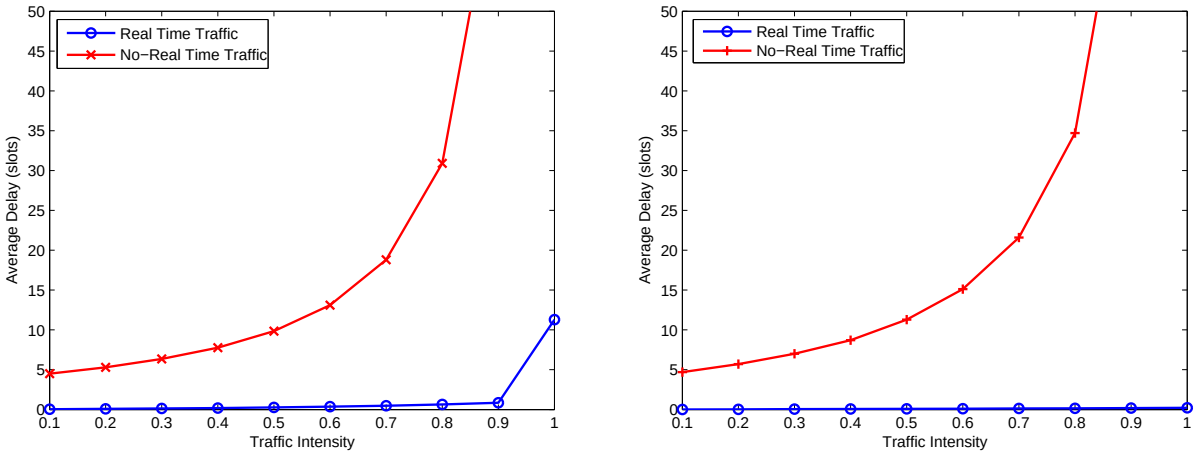
To analyze the behavior of the IEEE 802.16 MAC layer protocol regarding the delay caused by the uplink scheduling mechanism, this section presents some numeric results obtained with the proposed analytical model. The closed formulas presented in the previous section, were implemented in the MATLAB [15] software and will be used to obtain the results presented in this section. Moreover, through the simulation tool NS-2 (Network Simulator 2) [16], the proposed analytical model was also evaluated.

A. Analysis of the Data Allocation Phase

The real and non-real time messages delay is calculated in two different scenarios, where it is possible to compare the influence of a high load of flows with distinct priorities. These different classes of flows can be mapped in the four types of services offered by the IEEE 802.16 standard. This mapping is given as follows: the real-time class (higher priority) represents UGS and rtPS services while the non-real-time class represents nrtPS and BE services. The difference between these scenarios is that, in *Scenario I* there are more stations transmitting real-time traffic (70%UGS and 30%BE), while in *Scenario II* the classes of lower priority prevail (30%UGS and 70%BE).

The real and non-real time traffic average message delay are examined in Scenarios I and II, as a function of intensity $\rho_1 = \lambda_1 \nu_1$ and $\rho_2 = \lambda_2 \nu_2$, where $\rho = \rho_1 + \rho_2$, using the parameters in Table I.

Figures 3(a) and 3(b) illustrate the average waiting time in queue for each priority class as a function of the traffic offered in the channel. It is observed that, the wait time for high priority traffic (real-time) is much smaller in comparison to lower priority, even in Scenario II where a larger probability



(a) Scenario I

(b) Scenario II

Fig. 3. The Real and Non-Real messages delay.

of low priority traffics exists. The model in able to differentiate between classes of traffic efficiently, guaranteeing smaller waiting time in queue for higher priority messages. With that, the conformity of the model with the IEEE 802.16 standard can be demonstrated for the data messages.

B. Analysis of Bandwidth Request

In this section, the relative effectiveness of the bandwidth mechanism is investigated with the data traffic. In all scenarios presented in this section the non-real-time traffic load was fixed in $\rho_2 = 0.5$, so that the results were not influenced by the variation of the data load. Because the BE service schedule was used, the stations request bandwidth from the BS by sending bandwidth requests through contention.

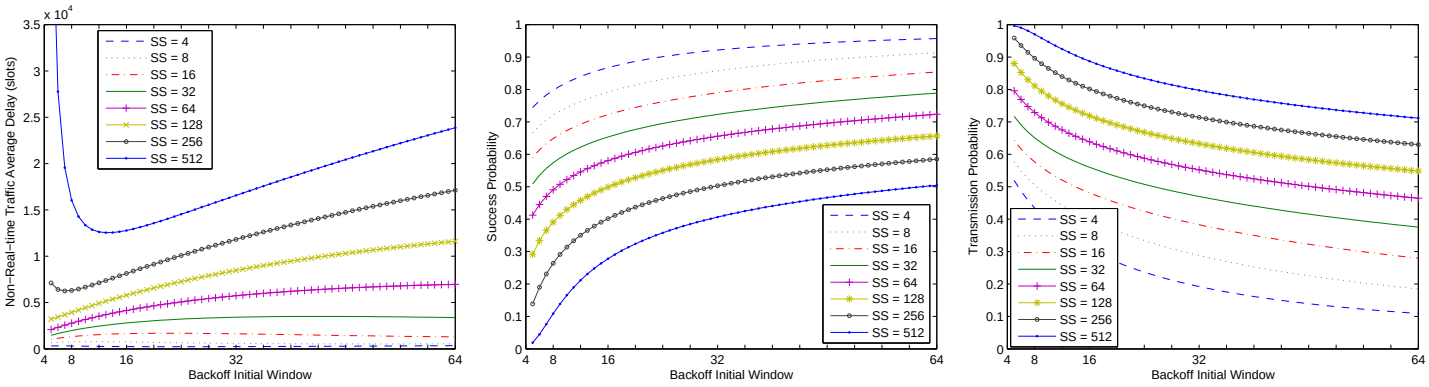
Figure 4(a) evaluates the impact of the initial backoff contention window in the performance, in terms of the bandwidth request messages average delay, considering the parameters $\{m = 6, R = 10, \mu = 8\}$. This figure shows the average delay when the minimum backoff window increases from 4 up to 64. Without taking into account the number of stations, in other words, by maintaining constant the number of stations requesting bandwidth, and increasing the initial backoff window size, the average message delay decreases to a minimum value, but starts increasing again, especially for a large number of stations. This behavior can be explained with the assistance of Figures 4(b) and 4(c). In Figure 4(b), it is noted that increasing the initial window size reduces message collisions, for any number of stations contending for the media and in Figure 4(c) it is noted that the transmission probability of a bandwidth request message depends on the initial backoff window and the number of stations in the network. Therefore, for larger contention windows, the station's probability of transmitting decreases, thus increasing the bandwidth request message delay due to longer waits for a transmission opportunity at the station.

Thus, when the number of stations contending for the media is large and the initial backoff window is small the probability of a collision happening is higher, which affects the message delay in a negative way. On the other hand, higher window sizes reduce the collision probability at the beginning of the contention process, making the delay decline. However, large initial windows increases the number of empty slots, making the delay increase again, due to the number of idle slots caused by the larger initial backoff window.

Figure 5(a) shows the bandwidth request messages average delay, considering the parameters $\{m = 6, R = 10, \mu = 8\}$, when the number of stations increase. The delay also grows due to the increase in the number of experienced collisions for the bandwidth request messages. Besides, the figure shows that the system delay is affected by the size of the initial backoff window. Small initial backoff windows provide smaller delays up to 220 stations, when the delay starts to grow quickly. This behavior can be explained by the fact that even if the request messages collide more, the delay for new transmission attempts is smaller, given that the backoff window size doubles after a collision happens. However, when increasing the number of stations, the bandwidth request messages suffer a lot with collisions and it takes a long time to converge toward windows where chances of collisions are minimized.

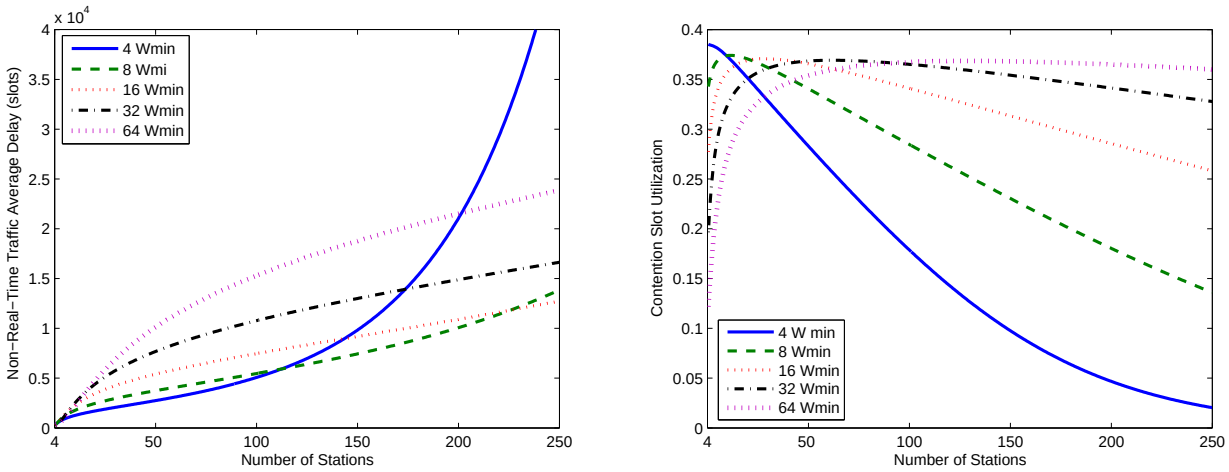
Figure 5(b) indicates the contention slot utilization for different values of the initial backoff window, considering the parameters $\{m = 6, R = 10, \mu = 8\}$. For each size of the initial backoff window there is an amount of stations that maximizes the slot contention use. When the number of stations and the size of the initial backoff window is small, the use of the slot is high and falls drastically with the increase in the number of stations. That behavior is due to the increase in the number of collisions for the bandwidth request messages. On the other hand, when the number of stations is small and the size of the initial backoff window is large, the usage of the segment is small, caused by empty slots. As the number of stations increases the segment utilization also increases, due to an increase of bandwidth request messages.

Figures 5(a) and 5(b), shows that the values that maximize slot usage, do not incur in the smallest message delays. That inconsistency happens due to the fact that a larger window



(a) Non-Real-Time average traffic delay. (b) Probability of a successful transmission. (c) Message transmission probability.

Fig. 4. Impact of the initial backoff contention window size in the bandwidth request messages performance



(a) Non-Real-Time Traffic Average Delay. (b) Contention Slot Utilization.

Fig. 5. Influence of the SS number performance analysis of the request bandwidth messages.

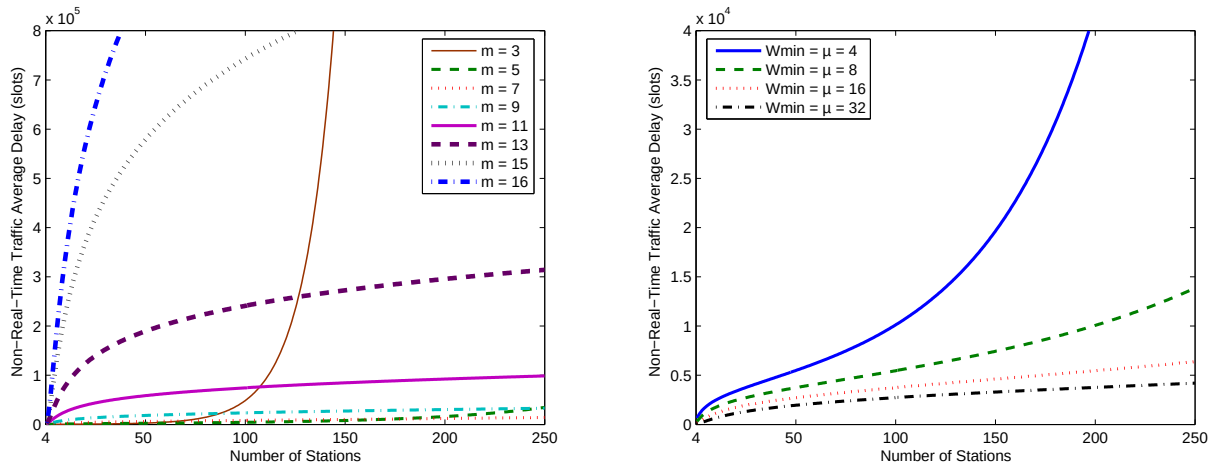
offers a larger penalty for message transmissions. A station that has a small backoff window size, and loses the dispute due to a collision, will need to wait fewer slots until its next attempt. That does not happen with stations that possess larger windows.

Figure 6(a) shows the non-real time message delay for different values of maximum backoff window, considering the parameters $\{W_{min} = 8, \mu = 8, R = 13, 11, 9, 7, 5, 3, 0\}$. When the number of stations increases from 4 to 512, the average delay also grows due to the increase in the number of collisions of bandwidth request messages. For small maximum backoff window sizes, the message delay is significantly lower, however, for maximum backoff window equal to 8 with more than 128 stations, the contention resolution process cannot solve the conflict. The big difference between the larger delays, offered by the largest backoff windows and the smaller delays, is due to the waiting time for a new transmission opportunity after a collision. Since the backoff window doubles its size after each collision, not limiting the maximum window size, implies significant increases in message delays. In this figure it is possible to see that the maximum window value that minimizes the delay, for up to 320 stations, is given by 32 slots, with m=5 and R=11.

The non-real-time total average delay for different contention periods, considering the parameters $\{m = 6, R = 10\}$, is presented in Figure 6(b). The initial backoff window is the same size as the contention period. When the number of stations grow, the message delay also grows due to collisions of bandwidth request message. The delay is influenced by the contention period, where greater contention periods offer smaller delays. However, increasing the contention period, implies diminishing the slots for data transmissions. Therefore, choosing smaller contention periods is desirable, given a good compromise between data rate and delay.

C. Model Validation

NS-2 (*Network Simulator*) [16] simulation tool was used to validate the model presented in this work. The NS-2 module for the MAC layer presented in [17] was used, so that it could be possible to simulate the IEEE 802.16 standard. There were 10 simulation runs for each scenario and the results presented in this section are the average and confidence interval with a



(a) Non-real-time traffic total average delay for different minimum backoff window sizes.

(b) Non-real-time traffic total average delay for different contention periods.

Fig. 6. Influence of backoff window size and the contention period on the bandwidth request messages total average delay.

confidence level of 95%.

Parameter	**Values**
Channel Bandwidth	40 Mbps
Frame Time	5 ms
Slot Size	250 bytes
Slot Time	0.05 ms
Backoff Initial Window	8 slots
Backoff Maximum Window	64 slots
Number of attempts in the maximum stage	10
Number of contention slots for the UL-MAP	8
Messages size	1 segment
Uplink frame size	50 slots
Duplexing Technique	TDD

TABLE I

SIMULATION PARAMETERS.

The objective of this simulation experiments is to analyze the behavior of the analytical model in a network with ideal channel conditions, in other words, without losses or messages corruption. The simulation scenario consists of a BS with stations uniformly distributed around it. This scenario does not have the intention of being representative for operational networks. The objective is to analyze the media access mechanism and the allocation of slots for different traffic intensities and number of stations. CBR sources were used to simulate the traffic of the two types of service flows. This was necessary, to facilitate the analysis of the results obtained through the analytical model. In all the simulated scenarios the presence of an admission control mechanism is assumed so that the results are not influenced by an excessive number of connections in the network. To avoid that the scheduling mechanism in the SSs interferes with the evaluation of the scheduling mechanism in the BS, each SS just generates a single data flow. The network configuration parameters can be found in Table I.

The real-time message delay modeling is validated using a simulation scenario composed by 1 BS and a number of SSs that varies from 1 to 84. The stations generate flows toward the uplink, with a data rate of 64 kbps and mapped for the UGS service. The grants interval is 10 ms because, in agreement with the IEEE 802.16 standard, the grants allocation interval at the BS and the packet generation interval at the SS application layer should be the same for that service.

Figure 7(a) presents the real-time messages average delay, obtained by simulation and through the proposed analytical model for a normalized traffic load varying from 0,01 to 1. The UGS traffic delay was not affected by the increase in the offered load, generated by increasing the number of stations. That indicates that the model can adapt to the UGS service delay requirements.

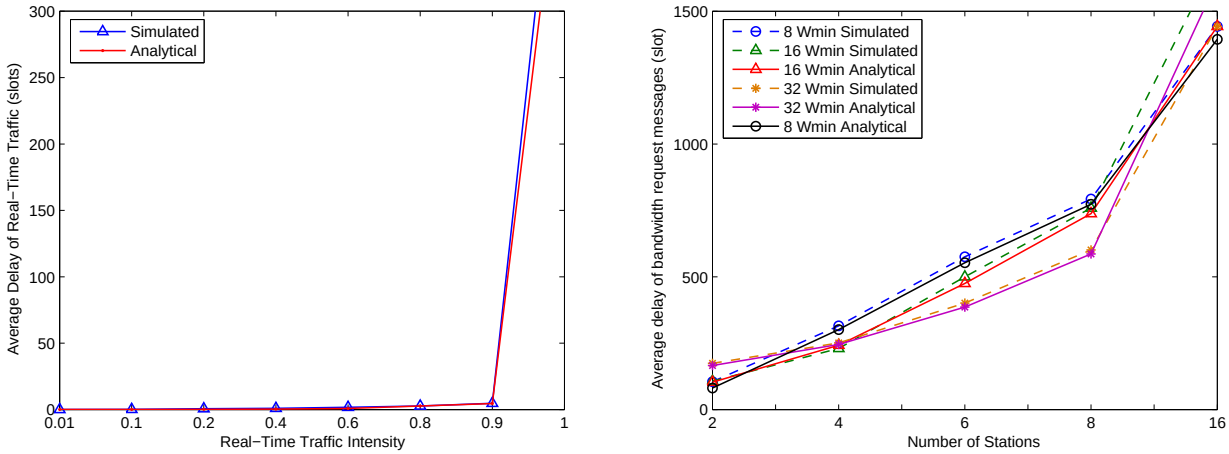
The model for the bandwidth request message delay is validated using a simulation scenario with 1 BS and a number of SSs varying from 2 to 16. The stations generate flows toward the uplink with a data rate of 200 kbps and mapped for the BE service.

Figure 7(b) shows the average delay for the request messages when the initial backoff window in equal to 8, 16 and 32 slots. Clearly the best time delay value is obtained for smaller values of W_{min} when the number of stations is small. On the other hand, when the number of stations increases, a smaller delay is reached with larger values of W_{min}.

V. CONCLUSION

Performance analysis through analytical modelling constitutes a technique of fundamental importance inside of the performance evaluation process. Existing literature investigates the impact of scheduling mechanisms in the performance of networks such as the ones studied here. However in previous work it is possible to find evaluations through simulation-based approaches or through analytical models, but with alterations of the IEEE 802.16 MAC layer protocol.

In this context, this work proposes an analytical model, applying queuing theory and Markov chains, to represent the behavior of the IEEE 802.16 MAC layer protocol in terms of end-to-end delay for real and non-real time traffics in bro-



(a) Real-Time-Traffic.

(b) Non-Real-Time Traffic with $W_{min} = 8$.

Fig. 7. Average Delay Analytical and Simulated

adband access wireless networks operating under the existing IEEE 802.16 standard. The performance of these networks can be evaluated, under the metrics of total data message delay and bandwidth request message delay, thus providing, a first step toward the specification of a complete model for schedule mechanisms for wireless metropolitan networks. Moreover, the proposed model allows the performance analysis of bandwidth request and data messages, whose resources were allocated through a contention process or through pre-established grants respectively.

The results obtained through the analytical modelling, verified the characteristics of the IEEE 802.16 MAC layer protocol. It is possible to see how, even for a higher load of non-real-time traffic, the proposed model is able to efficiently differentiate between classes of traffic, guaranteeing lower queue waiting times for higher priority messages, as it was defined in the standard. During the non-real-time messages analysis, it was possible to see how the maximum initial contention window and contention period highly influence the total average delay for these messages, given that the delay imposed by the bandwidth request messages is preponderant over the data messages total average delay. The process of contending for the media was also studied for scenarios with different number of stations and backoff parameters. In this case, for higher contention window sizes the probability of a station transmitting a message decreases, increasing the waiting time for a transmission opportunity and increasing consequently the bandwidth request message delay. Moreover, results show that a higher slot utilization does not mean smaller delay for the non-real-time traffic. Finally, the proposed analytical model was validated through simulation results, indicating the viability of using the analytical model for what it was intended.

As perspectives for future work, it is possible to carry on performance evaluations of the IEEE 802.16 standard through other important performance metrics, such as: throughput, jitter, loss probability and using other types of traffic sources.

REFERENCES

[1] IEEE 802.16, *IEEE Standard for Local and Metropolitan Area Networks - Part 16: Air Interface for Fixed Broadband Wireless Access Systems*, IEEE Std. 802.16, 2004.

[2] L. Kleinrock, *Queueing Systems, Vol. 2: Computer Applications*. New York, NY: Addison-Wesley, 1976.

[3] C. Cicconetti, C. Eklund, L. Lenzini, and E. Mingozzi, "Quality of service support in IEEE 802.16 networks," *IEEE Network*, vol. 20, pp. 50–55, Mar. 2006.

[4] C. Cicconetti, A. Erta, L. Lenzini, and E. Mingozzi, "Performance Evaluation of the IEEE 802.16 MAC for QoS Support," *IEEE Trans. Mob. Comp.*, vol. 6, pp. 26–38, 2007.

[5] K. Wongthavarawat and A. Ganz, "IEEE 802.16 Based Last Mile Broadband Wireless Military Networks with Quality of Service Support," *IEEE MILCOM*, vol. 2, pp. 779–784, 2003.

[6] D.-H. Cho, J.-H. Song, M.-S. Kim, and K.-J. Han, "Performance Analysis of the IEEE 802.16 Wireless Metropolitan Area Network," *DFMA*, pp. 130–137, 2005.

[7] S.-M. Oh and J.-H. Kim, "The Analysis of the Optimal Contention Period for Broadband Wireless Access Network," *PerCom*, pp. 215–219, 2005.

[8] N. O. O. Gusak and K. Sohraby, "Performance evaluation of the 802.16 medium access control layer," *Lecture Notes on Computer Science*, vol. 3280, pp. 228–237, 2004.

[9] D. Staehle and R. Pries, "Comparative Study of the IEEE 802.16 Random Access Mechanisms," *NGMAST*, pp. 334–339, 2007.

[10] B. Bhandari, R. Kumar, and S.L.Maskara, "Uplink Performance of the IEEE 802.16 Medium Access Control (MAC) Layer Protocol," *IEEE ICPWC*, pp. 23–25, 2005.

[11] R. Iyengar, P. Iyer, and B. Sikdar, "Analysis of 802.16 based last mile wireless networks," *IEEE GLOBECOM*, vol. 5, pp. 1–5, 2005.

[12] G. Bianchi, "Performance Analysis of the IEEE 802.11 Distributed Coordination Function," *IEEE Journal on Selec. Areas in Comm.*, vol. 18, no. 3, pp. 535–547, 2000.

[13] B. D. C. D. I. Choi and D. K. Sung, "Performance Analysis of Priority Leaky Bucket Scheme with Queue-Length-Threshold Schedule Policy," *IEE Proceedings Communications*, no. 145, pp. 395–401, 1998.

[14] L. Kleinrock, *Queueing Systems, Vol. 1: Theory*. New York, NY: Addison-Wesley, 1975.

[15] MATLAB. [Online]. Available: ⟨http://www.mathworks.com/⟩. Acessoem:2deset.2009

[16] The Network Simulator - ns-2. [Online]. Available: ⟨http://www.isi.edu/nsnam/ns/⟩.Acessoem:2deset.2009

[17] J. F. Borin and N. L. S. da Fonseca, "Um Módulo para Simulação de Redes WiMAX no Simulador NS-2," *VII Workshop de Desempenho de Sistemas Computacionais e de Comunicação, Anais do Congresso da Sociedade Brasileira de Computação*, pp. 1–15, 2008.