Upper Bound on Overall Delay in Wireless Broadband Networks with Non Real-Time Traffic

Sergey Andreev[1], Zsolt Saffer[2],
Andrey Turlikov[3], and Alexey Vinel[4]

[1] Tampere University of Technology (TUT), Finland
sergey.andreev@tut.fi
[2] Budapest University of Technology and Economics(BUTE), Hungary
safferzs@hit.bme.hu
[3] State University of Aerospace Instrumentation (SUAI), Russia
turlikov@vu.spb.ru
[4] Saint-Petersburg Institute for Informatics and Automation
Russian Academy of Sciences (SPIIRAS), Russia
vinel@ieee.org

Abstract. In this paper we consider the non real-time traffic in IEEE 802.16-based wireless broadband networks with contention-based bandwidth reservation mechanism. We introduce a new system model and establish an upper bound on the overall data packet delay. The model enables symmetric Poisson arrival flows and accounts for both the reservation and the scheduling delay components. The analytical result is verified by simulation.

Keywords: IEEE 802.16, queueing system, Markov chain, overall delay, contention-based request mechanism.

1 Introduction and Background

IEEE 802.16 telecommunication protocol [1] defined by the respective networking standard specifies a high data rate wireless broadband network with inherent support for various multimedia applications. Media access control (MAC) layer of IEEE 802.16 provides unified service for the set of physical (PHY) layer profiles, each of which corresponds to a specific operation environment. Currently we observe the proliferation of IEEE 802.16-based networks due to their relatively low cost, wide coverage and MAC mechanisms supporting a variety of quality of service (QoS) requirements.

Performance evaluation of IEEE 802.16 QoS mechanisms is addressed by numerous research papers. In particular, the so-called bandwidth reservation stage is often considered, at which a network user can reserve a portion of the channel resources. A general description of the different reservation techniques can be found in [2]. IEEE 802.16 protocol allows the usage of random multiple access

K. Al-Begain, D. Fiems, and W. Knottenbelt (Eds.): ASMTA 2010, LNCS 6148, pp. 262–276, 2010.

(RMA) for bandwidth requesting and it specifies the truncated binary exponential backoff (BEB) algorithm as means of collision resolution between the requests.

Asymptotic behavior of the BEB algorithm has been thoroughly investigated in the scientific literature. In [3] the BEB algorithm was shown to be unstable in the infinitely-many user model. By contrast, in [4] the BEB algorithm was demonstrated to be stable for sufficiently small arrival rates and finitely-many user model, even for the high number of users. Infinitely-many user model is known to highlight the limiting performance metrics of the algorithm, whereas finitely-many user model provides insight to the practical applicability of the algorithm. Finally, the operation of the BEB algorithm in the saturation conditions, where every network user always has pending data packets, was investigated by means of Markov models in [5] and [6].

Together with the separate analysis of the BEB collision resolution algorithm itself, its proper usage in the framework of IEEE 802.16 system is of interest. According to IEEE 802.16 protocol the BEB algorithm works with broadcast and multicast polling mechanisms (see [7] for details). The performance evaluation of broadcast polling was studied in [8]. Several important BEB application scenarios for the delay-sensitive traffic were discussed in [9].

The overall packet delay is strongly influenced by the choice of an appropriate bandwidth reservation mechanism. In [10] an efficient RMA algorithm is proposed, which may serve as an alternative to the standardized BEB algorithm at the reservation stage. IEEE 802.16 imposes no limitations on the methods for processing the bandwidth requests from the network users. Consequently, numerous scheduling algorithms were proposed.

For instance, in [11] a prioritized scheme for the request processing is developed together with the dynamic on-demand channel resource allocation. The performance of the proposed scheduling is also analyzed. A novel reservation algorithm is considered by [12], for which the corresponding analytical model is detailed. The model allows the evaluation of the reservation delay, but the scheduling delay is not addressed. Finally, in [13] an approach to estimate the overall packet delay is demonstrated. However, the scheduler-independent results there are approximations and thus they give only a rough delay estimate.

Therefore, we conclude that there is a lack of adequate models for the overall packet delay evaluation, including both reservation and scheduling delay. In our previous work [14] we gave an approximation for the overall packet delay. As a continuation of it in this paper we propose an analytical model that provides an upper bound on the overall data packet delay in IEEE 802.16 network.

2 IEEE 802.16 Short Summary

IEEE 802.16 standard specifies both PHY and MAC layers and provides dynamic resource allocation via bandwidth requesting and scheduling. Two operation modes are supported, where the point-to-multipoint mode is mandatory and the mesh mode is optional. MAC structure is composed of the three hierarchical sub-layers.

At the convergence sub-layer IP, ATM and Ethernet traffic is processed uniformly. At the common part sub-layer five different QoS profiles are defined. Various traffic flows with respective QoS requirements are mapped onto these profiles. According to the MAC specification the data packets may vary in size, subject to the proper aggregation/fragmentation. At the privacy sublayer the data encryption service is provided, as well as some additional cryptographic mechanisms.

The baseline PHY technology of IEEE 802.16 is orthogonal frequency division multiplexing (OFDM). Two OFDM-based schemes are defined: plain and OFD multiple access (OFDMA). Both schemes support adaptive modulation and coding to ensure reliable transmission under multipath propagation and over long distances. The growing number of IEEE 802.16 implementations are OFDMA-based, as OFDMA results in the higher spectral efficiency. However, consideration of OFDMA scheme is complicated due to the higher number of parameters and therefore we restrict our further explorations to the plain OFDM scheme.

The core IEEE 802.16 architecture comprises a base station (BS) and a set of subscriber stations (SSs) in its vicinity (see Figure 1). BS performs the polling of the SSs and manages the scheduling of SSs transmissions ensuring that the QoS guarantees of each data flow at each SS are satisfied. The BS and the SSs exchange packets via disjoint communication channels. In the downlink channel the BS broadcasts data to the SSs, whereas in the uplink channel the transmissions from the SSs are multiplexed.

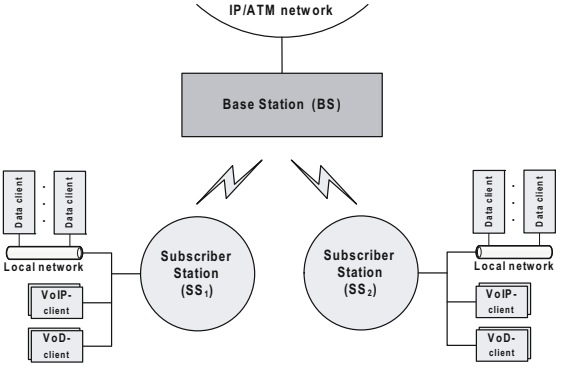

Fig. 1. Core IEEE 802.16 architecture

IEEE 802.16 provides two duplexing modes for the aforementioned downlink and uplink channels. In the time division duplex (TDD) mode a time frame is divided into downlink and uplink sub-frames, respectively. The simplified TDD frame structure is shown in Figure 2. In the frequency division duplex (FDD) mode the channel frequency range is divided into non-overlapping sub-ranges to avoid cross-interference.

As mentioned above, the BS broadcasts information to the wirelessly connected SSs. Together with the data packets, BS also sends relevant scheduling

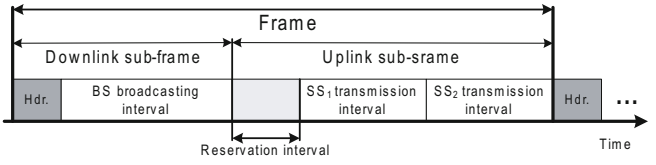

Fig. 2. Simplified TDD frame structure

information for both downlink and uplink channels. The uplink sub-frame schedule is incorporated into the UL-MAP (uplink map) management packet of the downlink sub-frame and is used by the SSs to determine the start time of their transmission in the uplink sub-frame. In order to enable the SSs to indicate their bandwidth needs to the BS, the so-called reservation interval, a portion of channel resources, is provided. The SSs are allowed to send their bandwidth requests during this interval. These requests are processed in the course of the scheduling.

There is a set of bandwidth requesting mechanisms at the reservation stage. Unicast polling is a contention-free mechanism, according to which the BS provides each SS with one transmission opportunity in a number of frames. Once provided, the transmission opportunity is used by the SS to send its bandwidth request. By contrast, broadcast and multicast polling are contention-based mechanisms. When broadcast polling is enabled the BS provides a number of transmission opportunities and each SS chooses one of them randomly. In case of multicast polling the SSs are grouped and broadcast polling is applied individually to each group. Simultaneous request transmissions may arise, if two or more SSs choose the same transmission opportunity to send their requests. Such request collisions are subject to the subsequent resolution by the BEB algorithm. Piggybacking feature allows an SS to append its bandwidth request to the transmitted data packet, when a connection to the BS is established.

As discussed previously, IEEE 802.16 successfully manages various multimedia connections. It is equally suitable for both high data rate (VoIP, audio and video) and low data rate (web) applications. The protocol supports bursty data flows and delay-sensitive traffic. In order to ensure the satisfaction of the QoS requirements for all these applications IEEE 802.16 standard introduces five QoS profiles. In particular, each profile specifies the type of a bandwidth requesting mechanism (contention-free and/or contention-based) to be used. Summarizing, a data flow with a dedicated identifier (ID) is mapped onto one of the following QoS profiles:

1. Unsolicited grant service (UGS). Used for real-time data sources with constant bit-rate (VoIP traffic without silence suppression). Uplink channel resource is granted periodically without explicit reservation.
2. Real-time polling service (rtPS). Used for real-time data sources with variable bit-rate (MPEG traffic). Uplink channel reservation is organized via unicast polling.

3. Extended real-time polling service (ertPS). Used for real-time data sources with variable bit-rate, which require more strict delay and throughput guarantees (VoIP traffic with silence suppression). This profile is introduced in one of the latest versions, IEEE 802.16e-2005 [15]. Uplink channel reservation is performed via unicast, multicast or broadcast polling.
4. Non real-time polling service (nrtPS). Used for non real-time data sources with variable packet length (FTP traffic). The allowed uplink channel reservation mechanisms are unicast, multicast or broadcast polling.
5. Best effort (BE). Used for non real-time data sources, which do not require delay and throughput guarantees (HTTP traffic). This profile utilizes the remaining bandwidth after scheduling all the above profiles. Multicast or broadcast polling can be used for uplink channel reservation.

Remember that all uplink transmissions are controlled by the BS scheduler. After a new data flow is mapped onto a particular QoS profile (UGS, rtPS, ertPS, nrtPS or BE) the SS proceeds with the uplink channel reservation by sending the corresponding bandwidth request. The BS sends back an UL-MAP management packet in the downlink sub-frame, which indicates the portion of the uplink sub-channel reserved for sending data packets.

The above summary implies that contention-based polling is the most widespread reservation mechanism in IEEE 802.16. Moreover, it is more difficult to analyze it due to its randomized nature in comparison to the analysis of the contention-free mechanism [16], [17]. Below we formulate a set of assumptions and detail the joint model to account for both the reservation and the scheduling stages.

3 System Model

In this section we describe the detailed model of IEEE 802.16-based network, which is used to evaluate the delay at both the reservation and the scheduling stages.

We consider the system that comprises a BS and M SSs, in which we focus only on the uplink transmissions. The BS is in the transmission range of all its SSs and all the SSs are in the reception range of the BS. In order to make the further analysis tractable we impose the following restrictions on the system operation according to IEEE 802.16 protocol description:

Restriction 1. The system operates in the point-to-multipoint mode.
Restriction 2. The time division duplex mode and the plain OFDM PHY scheme are used.
Restriction 3. The delay analysis is conducted for nrtPS QoS profile only, but both nrtPS and BE QoS profiles are considered.
Restriction 4. Only contention-based polling schemes are considered. We concentrate on the broadcast polling.

The system operation time is divided into frames and T_{frame} denotes the frame duration. The consecutive frames are indexed by integer nonnegative numbers,

$t = 0, 1, \ldots$. The duration of the packet transmission is τ. The packets arriving to SS i are also referred to as i-packets. At each SS the packet arrival process is Poisson. For simplicity we consider only symmetric arrival flows. Hence at each SS the arrival rate is the same, λ. Thus the overall arrival rate is $\Lambda = \lambda M$. The duration of each contention-based transmission opportunity is α. Moreover the reservation interval of each frame comprises exactly K contention-based transmission opportunities. A bandwidth request is issued by the i-th SS whenever at least one new data packet arrives, of which the BS should be notified. The request contains the information about all the newly arrived packets since the last request sending. If a packet arrives to an empty outgoing buffer of SS i during the reservation interval the SS must wait with sending the bandwidth request for this packet until the next reservation interval.

Additionally, below we introduce a set of assumptions to shape the system model. As such, we use the modified classical multiple access model, which is known from the substantial literature on multiple access techniques and applications, e.g. [18] and [19]. It is often addressed to compare various multiple-access protocols uniformly and has proved its usefulness over passing years.

1. The system
   - The number of contention-based transmission opportunities, K, is constant throughout the system operation.
   - The piggybacking is not used.
2. The BS
   - The BS maintains an individual grant buffer for each SS.
   - The individual BS buffers of the SSs have infinite capacity.
3. The SSs
   - Each SS is supplied with a infinite buffer to store data packets.
   - Each SS maintains exactly one active nrtPS connection.
   - Each SS can transmit exactly one packet in each uplink sub-frame.
4. The channel
   - The channel propagation time is negligible.
   - The uplink channel is noise-free. Consequently, if an SS transmits the BS always receives successfully. The downlink channel is also noise-free. Thus, all the SSs successfully receive the schedule of their transmissions.
   - In each contention-based transmission opportunity only one of the following events may arise at the same time: a single SS transmits its bandwidth request (SUCCESS), none of the SSs transmit (EMPTY), two or more SSs transmit their request simultaneously (COLLISION).
5. The feedback
   - The feedback for each SS about the success/failure of its own bandwidth request transmission (SUCCESS or NON-SUCCESS) is available. This feedback is necessary for the BEB algorithm operation.
   - The notification about the success of the bandwidth request transmissions is provided by the BS at the beginning of the following frame, that is, once in K transmission opportunities.

The BS uses the individual buffer of SS i to store the information about the number and the order of the i-packets (see Figure 3). At the end of each contention-based transmission opportunity the BS process a successfully received request, if any. The information about the newly arrived i-packets is extracted and placed into the corresponding BS buffer. Instead of each i-packet consideration it is equivalent to consider a grant assigned to it. These grants are placed into the individual BS grant buffer of SS i in the order of their extraction from the bandwidth request. This guarantees the first-come-first-served service.

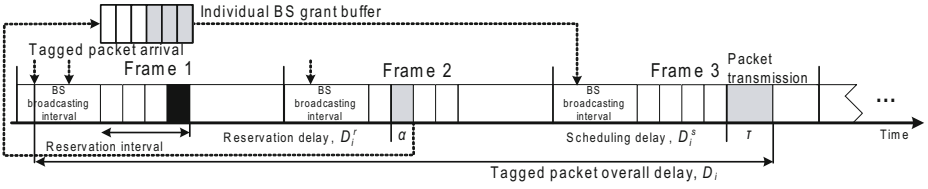

Fig. 3. An example of the request processing procedure

The BS processes the grants from the i-th buffer one by one until the buffer empties. During a frame only one grant can be processed from each grant buffer. When a grant is processed the BS forwards the scheduling information to the corresponding SS in the next frame for the uplink transmission of the corresponding packet. In case the i-th BS grant buffer was empty upon the reception of the new bandwidth request from SS i, the BS starts the service of the grants placed in the buffer immediately. Thus, the i-packet corresponding to the first i-grant will be included into the uplink schedule in the next frame.

When one or more SSs has empty BS grant buffer the BS utilizes the unused uplink transmission capacity to schedule BE packets. Such a behavior allows for avoiding the channel resource waste if individual BS buffer gets empty and therefore it results in a more efficient capacity utilization.

4 Overall Delay Analysis

In this section we conduct the evaluation of the overall data packet delay in the considered wireless broadband network. This delay includes both the reservation and the scheduling parts. We denote the durations of the downlink (DL) and the uplink (UL) sub-frames by T_{DL} and T_{UL}, respectively. Then for these durations it holds:

$$T_{UL} = T_{RI} + T_{UD}, \tag{1}$$

where T_{RI} is the duration of the reservation interval (RI) and T_{UD} is the maximum allowable duration of the UL sub-frame for sending the uplink data (UD).

Remember that according to the system model each frame comprises K contention-based transmission opportunities, which yields $T_{RI} = K\alpha$, where

α is the bandwidth request duration. Then we can rewrite the expression for T_{UD} as:

$$T_{UD} = T_{UL} - K\alpha. \tag{2}$$

Otherwise, accounting for the fact that each SS transmits at most one data packet per uplink sub-frame, we establish:

$$T_{UD} = M\tau. \tag{3}$$

Combining (1), (2) and (3) as well as assuming that the channel propagation time is negligible we obtain the following expression for the frame duration:

$$T_{frame} = T_{DL} + K\alpha + M\tau. \tag{4}$$

Let ρ denote the load at SS i. As an SS transmits at most one packet per frame, we obtain:

$$\rho = \lambda T_{frame} = \frac{\Lambda T_{frame}}{M}. \tag{5}$$

Clearly, the considered system is stable when $\rho < 1$ or $\Lambda < \frac{M}{T_{frame}}$, that is the number of arriving packets does not on average exceed the number of departing packets.

Consider the overall packet delay D_i for the i-th SS, which is a continuous random variable. This delay arises due to both queueing in the outgoing SS buffer during the reservation delay and queueing in the BS buffer during the scheduling delay. The overall packet delay is thus defined as the time interval from the moment the packet arrives into the system to the moment when its successful uplink transmission ends. Figure 3 illustrates the following components of the overall tagged packet delay:

$$D_i = D_i^r + \alpha + D_i^s + \tau, \tag{6}$$

where the components are defined as follows.

- D_i^r – reservation delay from the moment the packet arrives into the outgoing buffer of SS i to the start of the successful transmission of the corresponding bandwidth request in the reservation interval.
- α – time of the successful bandwidth request transmission, which equals the duration of the transmission opportunity.
- D_i^s – scheduling delay from the end of the successful bandwidth request transmission of the i-th SS to the start of the corresponding data packet transmission in the uplink sub-frame.
- τ – data packet transmission time.

The main assumption of the analysis is that the probability of the successful bandwidth request transmission in a reservation interval is constant. Let p_r denote this probability as it is independent of the SS index. Accounting for the fact that each SS has an individual BS buffer and an own, separate data packet transmission period in the uplink sub-frame, we conclude that the statistical

behavior of an SS is independent of that one for the other SSs. As such, to establish the overall packet delay of the tagged SS, it is enough to model its behavior separately from the rest of the system.

According to this we consider the system shown in Figure 3 from the point of view of the tagged SS i. For the sake of simplicity in the following description we omit the index i. We construct an embedded Markov chain [20] at the sequence of begin times of the consecutive reservation intervals. The state of the chain consists of the number of packets in the SS and BS buffers. More precisely, we assume that there are three buffers for the data packets (see Figure 4). The first buffer is the one at the tagged SS where the packet is queued during the reservation delay. After that the packet is immediately transferred to the virtual buffer at the beginning of the corresponding reservation interval. The virtual buffer accounts for the fact that a packet cannot be transmitted in the current frame, that is, experiences the delay of at least one frame. After this additional delay the packet enters the individual BS buffer of the tagged SS. There the packet is queued until the end of the scheduling delay. Finally, the packet is transmitted. Note that in this equivalent queueing system we implicitly assume that the transitions between the buffers and leaving the last buffer happen at the embedded epochs, i.e. somewhat earlier comparing to e.g. the BS processing at the end of the contention-based transmission opportunities.

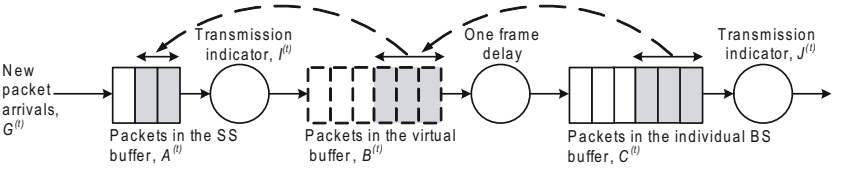

Fig. 4. Equivalent queueing system description

Let $\{A^{(t)}\}$, $\{B^{(t)}\}$ and $\{C^{(t)}\}$ denote the number of packets in the first buffer, in the virtual buffer and in the individual BS buffer at the embedded epoch in the t-th frame, respectively. The dynamics of the number of packets in the first SS buffer at the consecutive embedded time epochs in frame t and $t+1$ can be expressed by the following expression:

$$A^{(t+1)} = (A^{(t)} + G^{(t)})(1 - I^{(t)}), \tag{7}$$

where $G^{(t)}$ is the number of newly arriving packets, which enter the SS buffer during the interval between the t-th and $(t+1)$-th embedded epochs and $I^{(t)}$ is the discrete indicator function showing if the corresponding bandwidth request is transmitted successfully in the reservation interval of the t-th frame:

$$I^{(t)} = \begin{cases} 1 & \text{with probability } p_r, \\ 0 & \text{with probability } 1 - p_r. \end{cases} \tag{8}$$

The dynamics of the number of packets in the virtual buffer $\{B^{(t)}\}$ could be described as follows:

$$B^{(t+1)} = (A^{(t)} + G^{(t)})I^{(t)}. \tag{9}$$

Finally, the evolution of the number of packets in the individual BS buffer $\{C^{(t)}\}$ at the embedded time moments could be written as:

$$C^{(t+1)} = C^{(t)} - J^{(t)} + B^{(t)}, \tag{10}$$

where $J^{(t)}$ is the discrete indicator function showing if the packet is transmitted successfully in the uplink sub-frame of the t-th frame:

$$J^{(t)} = \begin{cases} 1, & \text{if } C^{(t)} > 0, \\ 0, & \text{if } C^{(t)} = 0. \end{cases} \tag{11}$$

We are interested in obtaining the steady-state expression for the mean number of packets in all the considered buffers. Let $E[A^j]$, $E[B^j]$ and $E[C^j]$ stand for the limiting j-th moments of the $\{A^{(t)}\}$, $\{B^{(t)}\}$ and $\{C^{(t)}\}$ random variables, for $j = 1, 2, \ldots$, i.e. $E[A^j] = \lim\limits_{t\to\infty} E[(A^{(t)})^j]$, $E[B^j] = \lim\limits_{t\to\infty} E[(B^{(t)})^j]$ and $E[C^j] = \lim\limits_{t\to\infty} E[(C^{(t)})^j]$. The ergodicity of the considered Markov chain ensures the existence of the limiting distributions of $\{A^{(t)}\}$, $\{B^{(t)}\}$ and $\{C^{(t)}\}$.

To determine the steady-state mean number of packets in all three buffers we derive relations for the above first two moments from the equations (7), (9) and (10). We average both parts of these expressions. Also we raise them to the second power and then take the mathematical expectation of both parts. Utilizing the mutual independence of $A^{(t)}$, $G^{(t)}$ and $I^{(t)}$ as well as the mutual independence of $B^{(t)}$ and $C^{(t)}$ the required mean quantities can be established as:

$$\begin{aligned} E[A] &= \rho\frac{1-p_r}{p_r}, \\ E[B] &= \rho, \\ E[C] &= \frac{2\rho - \rho^2\left(3 - \frac{2}{p_r}\right) - 2E[AJ]}{2(1-\rho)}. \end{aligned} \tag{12}$$

We note that the expression of $E[C]$ explicitly incorporates $E[AJ]$, since the $A^{(t)}$ and $J^{(t)}$ are dependent random variables. The exact calculation of this term is not easy and therefore here we replace it with zero. As such, we derive the upper bound on $E[C]$. The total number of packets at the embedded epoch is $E[A] + E[B] + E[C]$. Note that in the three buffers model every transition between the buffers occurs at an embedded epoch and only packet arrivals happen between these epochs. If the state of the Markov chain at an embedded epoch is given, the stochastic evolution of the the number of packets in the system repeats itself in the intervals between the consecutive embedded epochs. Thus, to get the number of packets in the system at an arbitrary epoch (Q^{rs}) it is enough to consider it at an arbitrary epoch in the intervals between two consecutive embedded epochs

with the length of T_{frame}. Hence $E[Q^{rs}]$ plus the number of arriving packets during the forward recurrence time of such an interval ($\lambda\frac{T_{frame}}{2}$) is exactly the total number of packets at the embedded epochs. This yields:

$$E[Q^{rs}] = E[A] + E[B] + E[C] - \frac{\rho}{2}. \tag{13}$$

Applying (13) in the Little's formula [21] we can obtain the upper bound on the sum of the reservation and the scheduling packet delays. However the virtual buffer accounts only for a part of the delay from scheduling a grant of a packet to the start of the uplink transmission of that packet. The rest of this delay for an i-packet is given as $\alpha K + (i-1)\tau$. Averaging over every possible $i = 1, \ldots, M$ yields $\alpha K + \tau\frac{M-1}{2}$. Furthermore as the embedded epoch happens at least by α time earlier as the end of the contention-based transmission opportunities the above delay part is upper bounded by $\alpha(K-1) + (i-1)\tau$. Taking this term into account in the scheduling packet delay, expressing the sum of the reservation and the scheduling packet delays by applying Little's formula and using (6) results in the upper bound on overall packet delay in the considered wireless broadband network as:

$$E[D] \leq \left(\frac{3}{2} + \frac{1-p_r}{p_r}\right) T_{frame} + \frac{\rho T_{frame}\,(2-p_r)}{2p_r(1-\rho)} + \alpha K + \tau\frac{M+1}{2}. \tag{14}$$

The probability of the successful bandwidth request transmission in a reservation interval p_r can be determined by means of a second Markov chain model, which uses the quantity p_t, which is defined as the probability of a transmission attempt of an SS.

Firstly, we briefly summarize the determination of the probability p_t, which is presented in [22]. At the reservation stage IEEE 802.16 users follow the rules of the BEB algorithm used for the collision resolution. The BEB algorithm operation is thoroughly investigated in [22]. According to [5] and [6] the consideration of the entire system could be reduced to the consideration of the tagged SS only. For a contention-based transmission opportunity the conditional collision probability, conditioning on the fact that the SS attempts the transmission (p_c) is introduced as:

$$p_c = 1 - (1-p_t)^{M-1}. \tag{15}$$

This probability may be established by:

$$p_t = \frac{2(1-2p_c)}{(1-2p_c)(W_0+K) + p_c W_0(1-(2p_c)^m)}, \tag{16}$$

where W_0 and m are the parameters of the BEB algorithm and they are termed as initial contention window and maximum backoff stage, respectively. Hence, the probabilities p_t and p_c can be determined by solving the system of two non-linear equations (15) and (16).

As stated above having the probability p_t a second Markov chain model can be set up for determination of p_r. This can be described analogously to its description in our previous work [14]. From the point of view of the bandwidth requesting each SS may reside in an active or an inactive state. Active SS participates in the contention process, i.e. it has at least one pending data packet, for which a successful bandwidth request has not yet been issued. Inactive SS does not initiate the reservation process as it has no packets, of which the BS has not yet been successfully informed. We introduce a Markov chain embedded at the sequence of the ends of the contention-based transmission opportunities. The state of this Markov chain $\{N^{(u)}\}$, for $u = 1, \ldots$, composes of the number of active SSs. In each frame the first packet arrives to an inactive SS with the probability $y = 1 - e^{-\lambda T_{frame}}$. After the first packet arrival the SS enters the active state, issues a new bandwidth request and starts the contention process, for which all the subsequent arrivals are irrelevant. According to these the transition probabilities among the $M + 1$ states of the chain can be written as:

$$p_{i,j} = \Pr\{N^{(t+1)} = j | N^{(t)} = i\} = \tag{17}$$

$$= \begin{cases} 0, & \text{if } j \leq i-2, \\ ip_t(1-p_t)^{i-1}(1-y)^{M-i+1}, & \text{if } j = i-1, \\ ip_t(1-p_t)^{i-1}(M-i+1)y(1-y)^{M-i} + & \\ +(1-ip_t(1-p_t)^{i-1})(1-y)^{M-i}, & \text{if } j = i, \\ ip_t(1-p_t)^{i-1}\binom{M-i+1}{j-i+1}y^{j-i+1}(1-y)^{M-j} + & \\ +(1-ip_t(1-p_t)^{i-1})\binom{M-i}{j-i}y^{j-i}(1-y)^{M-j}, & \text{if } j \geq i+1. \end{cases}$$

It may be shown that the considered Markov chain is finite and irreducible for $p_t, y > 0$ [23]. Therefore, its stationary probability distribution exists, which may be obtained by solving a linear system of $M + 1$ equations:

$$\begin{cases} P_j = \sum_{i=0}^{M} P_i p_{i,j} \quad \text{for } j = 0, 1, \ldots, M, \\ \sum_{i=0}^{M} P_i = 1. \end{cases} \tag{18}$$

We determine the joint probability at the end of a contention-based transmission opportunity that the number of active SSs is n ($n = 1, \ldots, M$) and the tagged SS is among them and the tagged SS has successful bandwidth request transmission. This probability is denoted by $s(n)$. Due to the symmetry of the model the probability that the tagged SS is among the i active SSs is given by $\frac{\binom{M-1}{n-1}}{\binom{M}{n}} = \frac{n}{M}$. Thus the $s(n)$ can be expressed as:

$$s(n) = \frac{n}{M} p_t (1-p_t)^{n-1}. \tag{19}$$

Let p_s denote of the successful bandwidth request transmission of the tagged SS at the end of a contention-based transmission opportunity. p_s can be calculated with the help of the stationary distribution $\{P_n\}_{n=\overline{0,M}}$ of the Markov chain as:

$$p_s = \sum_{n=0}^{M} s(n) P_n. \tag{20}$$

A bandwidth request transmission in a reservation interval can be successful in any of the K provided contention-based transmission opportunity. As these events exclude each other, p_r can be given by:

$$p_r = K p_s. \tag{21}$$

5 Numerical Results and Conclusion

In order to verify the adequacy of the model assumptions made during the performance analysis we developed a simplified IEEE 802.16 MAC simulator. It accounts for the restrictions of the system model and was previously used in [7], [16], [14] and [17]. According to [24] we set the typical simulation parameters and summarize them in Table 1.

Table 1. Typical simulation parameters

IEEE 802.16 parameter	Value
DL:UL proportion	60:40
PHY type	OFDM
Frame duration (T_{frame})	5 ms
Channel bandwidth	7 MHz
Contention-based transmission opportunity duration (α)	170 μs
Data packet length	4096 bits

The result of the verification for this typical parameter set is demonstrated in Figure 5, where curves show analytical results and symbols are obtained with simulation. The accuracy of the model depends on the overall arrival rate and some system parameters, such as p_r. Although we do not include results for different values and system parameters in this paper, we have shown through extensive simulations that the derived model is reasonably accurate for the realistic protocol settings. Therefore, it is a useful tool for the evaluation of the overall packet delay, as well as for fine-tuning the wireless system to control it.

In this paper we proposed an analytical model to estimate the overall data packet delay in IEEE 802.16 network. The model accounts for the delay at both the reservation and the scheduling stages. Several assumptions of the presented model can be relaxed and hence the analysis can be generalized in these directions. According to IEEE 802.16 protocol each SS may potentially establish **multiple connections** with the BS. The developed system model may be

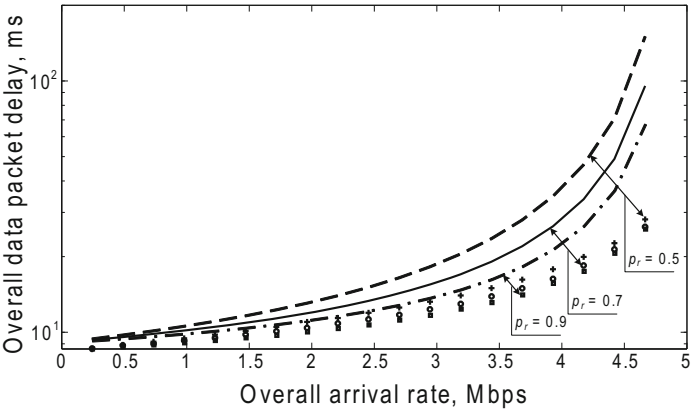

Fig. 5. Overall data packet delay in logarithmic scale for $M = 6$ and $K = 1$

generalized for this case by considering connections instead of SSs. The assumption about the noise-free uplink and downlink channels is also non-realistic. In practice the transmissions are always corrupted by the adverse wireless channel effects. The analytical model enables the extension for the case of the **noisy channel**. Finally, the proposed analytical model could be modified to account also for the **unicast polling** of the SSs incorporating the models of [16] and [17].

Acknowledgments

This work is supported by the Russian Foundation for Basic Research (projects # 10-08-01071-a and # 08-08-00403-a), as well as by the Branch of Nano- and Information Technologies of Russian Academy of Sciences (project 2.3)

References

1. IEEE 802.16-2009. IEEE Standard for Local and metropolitan area networks (May 2009)
2. Rubin, I.: Access-control disciplines for multi-access communication channels: reservation and TDMA schemes. IEEE Transactions on Information Theory 25(5), 516–536 (1979)
3. Aldous, D.: Ultimate instability of exponential back-off protocol for acknowledgment based transmission control of random access communication channels. IEEE Transactions on Information Theory 33(2), 219–223 (1987)
4. Goodman, J., Greenberg, A., Madras, N., March, P.: Stability of binary exponential backoff. Journal of the ACM 35(3), 579–602 (1988)
5. Bianchi, G.: Performance analysis of the IEEE 802.11 distributed coordination function. IEEE Journal on Selected Areas in Communications 18(3), 535–547 (2000)
6. Song, N., Kwak, B., Miller, L.: On the stability of exponential backoff. Journal of Research of the NIST 108(4), 289–297 (2003)

7. Andreev, S., Turlikov, A., Vinel, A.: Contention-based polling efficiency in broadband wireless networks. In: Proc. of the 15th International Conference on Analytical and Stochastic Modeling Techniques and Applications, pp. 295–309 (2008)
8. Lin, L., Jia, W., Lu, W.: Performance analysis of IEEE 802.16 multicast and broadcast polling based bandwidth request. In: Proc. of the IEEE Wireless Communications and Networking Conference, pp. 1854–1859 (2007)
9. Alanen, O.: Multicast polling and efficient VoIP connections in IEEE 802.16 networks. In: Proc. of the 10th ACM Symposium on Modeling, Analysis, and Simulation of Wireless and Mobile Systems, pp. 289–295 (2007)
10. Kobliakov, A., Turlikov, A., Vinel, A.: Distributed queue random multiple access algorithm for centralized data networks. In: Proc. of the 10th IEEE International Symposium on Consumer Electronics, pp. 290–295 (2006)
11. Cho, D., Song, J., Kim, M., Han, K.: Performance analysis of the IEEE 802.16 wireless metropolitan network. In: Proc. of the 1st International Conference on Distributed Frameworks for Multimedia Applications, pp. 130–136 (2005)
12. Moraes, L., Maciel, P.: Analysis and evaluation of a new MAC protocol for broadband wireless access. In: Proc. of the International Conference on Wireless Networks, Communications and Mobile Computing, vol. 1, pp. 107–112 (2005)
13. Iyengar, R., Iyer, P., Sikdar, B.: Delay analysis of 802.16 based last mile wireless networks. In: Proc. of the 48th IEEE Global Telecommunications Conference, vol. 5, pp. 3123–3127 (2005)
14. Andreev, S., Saffer, Z., Turlikov, A., Vinel, A.: Overall delay in IEEE 802.16 with contention-based random access. In: Proc. of the 16th International Conference on Analytical and Stochastic Modeling Techniques and Applications, pp. 89–102 (2009)
15. IEEE 802.16e-2005. Amendment to IEEE Standard for Local and Metropolitan Area Networks (February 2006)
16. Saffer, Z., Andreev, S.: Delay analysis of IEEE 802.16 wireless metropolitan area network. In: Proc. of the 15th International Conference on Telecommunications, pp. 1–5 (2008)
17. Andreev, S., Saffer, Z., Anisimov, A.: Overall delay analysis of IEEE 802.16 network. In: Proc. of the IEEE International Conference on Communications, pp. 1–6 (2009)
18. Bertsekas, D., Gallager, R.: Data Networks. Prentice-Hall, Englewood Cliffs (1992)
19. Rom, R., Sidi, M.: Multiple Access Protocols: Performance and Analysis. Springer, Heidelberg (1990)
20. Kleinrock, L.: Queueing Systems: Volume II - Computer Applications, New York (1976)
21. Kleinrock, L.: Queueing Systems: Volume I – Theory, New York (1975)
22. Andreev, S., Turlikov, A.: Binary exponential backoff algorithm analysis in the lossy system with frames. In: Proc. of the 12th International Symposium on Problems of Redundancy in Information and Control Systems, pp. 201–210 (2009)
23. Kleinrock, L., Lam, S.: Packet-switching in a multi-access broadcast channel: performance evaluation. IEEE Transactions on Communications 23(4), 410–423 (1975)
24. Sivchenko, D., Bayer, N., Xu, B., Rakocevic, V., Habermann, J.: Internet traffic performance in IEEE 802.16 networks. In: Proc. of the 12th European Wireless Conference, pp. 1–5 (2006)