Overall Delay in IEEE 802.16 with Contention-Based Random Access*

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

[1] State University of Aerospace
Instrumentation (SUAI), Russia
serge.andreev@gmail.com
turlikov@vu.spb.ru

[2] Department of Telecommunications,
Budapest University of Technology,
and Economics (BUTE), Hungary
safferzs@hit.bme.hu

[3] St.Petersburg Institute for Informatics
and Automation of RAS (SPIIRAS), Russia
vinel@ieee.org

Abstract. In this paper we address the overall message delay analysis of IEEE 802.16 wireless metropolitan area network with contention-based multiple access of bandwidth requests. The overall delay consists of the reservation and scheduling components. Broadcast polling is used for bandwidth reservation with binary exponential backoff (BEB) collision resolution protocol and a simple scheduling is applied at the base station. An analytical model is developed with Poisson arrival flow for the Non Real-Time Polling Service (nrtPS) class. The model enables asymmetric traffic flows, different message sizes at the subscriber stations and also allows for Best Effort (BE) service class. An approximation of the mean overall delay is established for the nrtPS service class. The analytical model is verified by means of simulation.

Keywords: IEEE 802.16, WMAN, performance evaluation, bandwidth reservation, contention-based multiple access, BEB, queueing model.

1 Introduction

IEEE 802.16 is a notorious specification, which is recommended for Wireless Metropolitan Area Networks (WMANs). The standard specifies an air interface for Broadband Wireless Access (BWA) [1]. It proposes a high-speed access system supporting multimedia services and an extensive *quality-of-service* (QoS) guarantee. In IEEE 802.16 protocol stack the Medium Access Control (MAC) layer supports multiple Physical (PHY) layer specifications, each of them covering different operational environments.

* This work is partially supported by the NAPA-WINE FP7-ICT (http://www.napa-wine.eu) and the OTKA K61709 projects.

K. Al-Begain, D. Fiems, and G. Horváth (Eds.): ASMTA 2009, LNCS 5513, pp. 89–102, 2009.

Many authors studied the performance of the various IEEE 802.16 features. In particular, the bandwidth requests mechanism to reserve a portion of the channel resources is frequently addressed. A detailed description of the reservation techniques and a general queueing model are given in the fundamental works [2] and [3]. The standard allows a *random multiple access* (RMA) reservation scheme and implements the truncated *binary exponential backoff* (BEB) protocol for the purposes of the collision resolution.

The asymptotic behavior of the BEB protocol was substantially addressed in the literature. In [4] it was shown that the BEB protocol is *unstable* in the infinitely-many users case. By contrast, [5] shows that the BEB is *stable* for any finite number of users, even if it is extremely large, and sufficiently low input rate. An exhaustive description of various analytical RMA models models may be found in [6] and [7]. The performance of the BEB algorithm in the framework of the reference RMA model ([8], [9]) is addressed in [10], which allowed a deeper insight into its operation. In the fundamental analysis of [11] an extremely useful Markovian model to analyze the performance of the BEB algorithm was first introduced.

Together with the analysis of the BEB itself, much attention is paid to its proper usage in IEEE 802.16 standard. It is known that the BEB algorithm may be adopted for both *broadcast* and *multicast* user polling. The efficiency of broadcast and multicast polling was extensively studied in [12] and [13]. Some practical aspects of the BEB application for the delay-sensitive traffic were considered in [14].

Considerable research effort is done also on overall performance aspects of the IEEE 802.16 system. For example in [15], [16] and [17] various frameworks are built and analyzed to guarantee a specified level of QoS. Furthermore, in [18] and [17] the overall system delay is estimated and verified. However none of these methods are dealing with overall delay in the context of contention-based random access.

In this paper we develop a first analytic approximation for the overall delay in the IEEE 802.16 system with broadcast polling.

The rest of the paper is structured as follows. Section II gives a brief overview of IEEE 802.16 MAC layer. In Section III we provide the description of the model and the notations. We conduct the overall delay analysis in Section IV. In Section V we verify the analytical results by means of simulation. Finally, we give our conclusion in Section VI.

2 Brief Overview of IEEE 802.16 MAC

IEEE 802.16 standard supports two operational modes: the mandatory *Point-to-MultiPoint* (PMP) and the optional mesh mode. In the centralized PMP architecture the *Base Station* (BS) is the main node. It is responsible for coordinating the communication process among the other nodes – *Subscriber Stations* (SSs). All communication among the SSs is directed through the BS and takes place on independent transmission channels of two types. In the *Downlink Channel* (DL) only the BS transmits data to the SSs, while in the *Uplink Channel* (UL) the

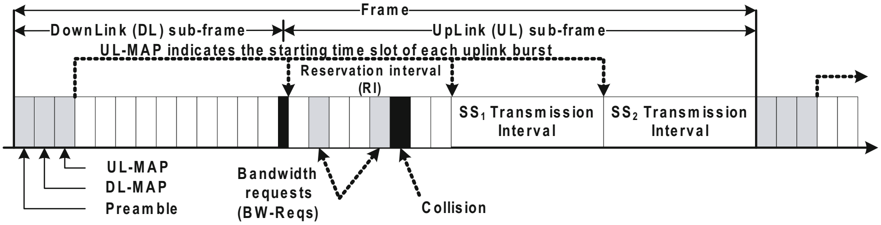
Frame
DownLink (DL) sub-frame
UpLink (UL) sub-frame
UL-MAP indicates the starting time slot of each uplink burst
Reservation interval (RI)
SS_1 Transmission Interval
SS_2 Transmission Interval
UL-MAP
DL-MAP
Preamble
Bandwidth requests (BW-Reqs)
Collision

Fig. 1. IEEE 802.16 MAC frame structure in TDD/TDMA mode

data is sent by the SSs to the BS. Hence, there is no multiple access on the DL channel, while the UL channel is shared among multiple SSs.

The standard provides two channel allocation schemes: *Frequency Division Duplexing* (FDD) and *Time Division Duplexing* (TDD). In FDD the the DL and the UL channels are assigned to the different frequencies, while in TDD both channels are assigned to the same frequency, and are differentiated by assigning different time intervals to them. In this case the time is divided into fixed-length frames, which consist of the DL and the UL sub-frames corresponding to the DL and the UL channels, respectively. The length of the sub-frames can be varied dynamically. The SSs access the UL channel by means of *Time-Division Multiple Access* (TDMA).

The MAC frame structure can be seen in Figure 1. In the DL sub-frame the BS broadcasts data to all the SSs, and each of them captures only those addressed to it. Besides the DL scheduling, the BS is also responsible for the UL scheduling. The BS determines the number of slots to be allocated for each SS in the next UL sub-frame. This information is broadcasted in the UL-MAP message in the beginning of each frame. After receiving the UL-MAP message, the SS transmits data in the next UL sub-frame using the time slots which are granted to it.

The SS can initiate bandwidth reservation by sending a *Bandwidth Request* (BW-Req) message in the *Reservation Interval* (RI) in the beginning of each UL sub-frame. The standard defines contention-free polling mechanism (unicast) and contention-based random access polling mechanisms (multicast or broadcast) for bandwidth reservation. The duration of the RI is not specified by the standard explicitly. In case of contention-based random access, the defined collision resolution mechanism is the truncated *Binary Exponential Backoff* (BEB) protocol. Additionally, IEEE 802.16 enables piggybacking for sending BW-Reqs attached to data packets.

3 Model and Notations

3.1 Restrictions of the Model

Our model describes the IEEE 802.16 MAC with the following limitations:

R.1: The operational mode is PMP.

R.2: TDD/TDMA channel allocation scheme is used.

R.3: Messages of nrtPS and BE service classes are allowed, however we consider only the performance of the nrtPS service class.

R.4: The bandwidth reservation mechanisms is the contention-based broadcast polling.

R.4: The uplink scheduler applies a simple scheduling (see in 3.2).

R.6: One connection per SS is allowed.

R.7: Piggybacking is not used.

3.2 General Model and Scheduling

There are 1 BS and N SSs in the system, which together comprise $N+1$ stations. In this model we consider only the uplink traffic of messages. Each SS has infinite buffer capacity to store the waiting messages. Messages transmitted by the SSs consist of a number of data packets.

A BW-Req sent by a SS i represents the request for all i-messages, which are accumulated in its outgoing buffer since its last successful BW-Req sending.

For each SS the BS maintains an individual buffer with infinite capacity. At the end of each polling slot the BS performs an immediate processing of the successfully received BW-Req, if any, and of non-empty individual BS buffers of SSs, at which scheduling arises.

If BW-Req is received from SS i, then BS immediately assigns an individual request to each data message represented by the received BW-Req and these requests are put into the corresponding individual BS buffer of SS i (according to their order taken from BW-Req). If the individual BS buffer of SS i is empty upon receiving a BW-Req from that SS, then after putting the individual requests into the individual BS buffer of that SS the i-message corresponding to the first individual request is immediately scheduled for transmission on UL in the next frame. The frame duration is T_f. As exactly one i-message can be transmitted in one frame the BS schedules the i-messages corresponding to the waiting individual requests periodically after each T_f time until the individual BS buffer of SS i becomes empty.

The contention-based random access and the scheduling is illustrated in Figure 2.

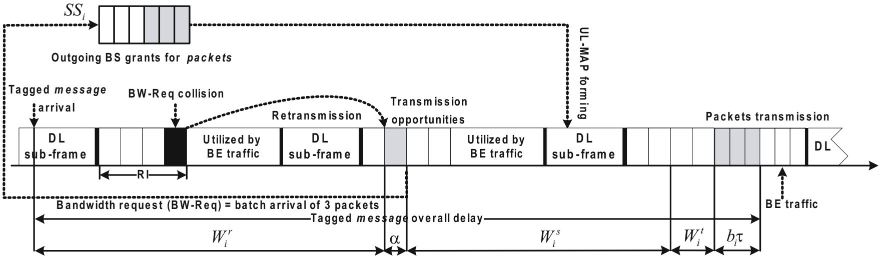

Fig. 2. Contention-based model and scheduling operation

This way one message will be scheduled from every non-empty individual BS buffers of SSs during processing a frame (its RI). This processing is repeated periodically in consecutive frames.

In the case when one or more SSs have no message of nrtPS service class to send on uplink, the system is allowed to utilize the unused uplink transmission capacity for uplink transmission of BE messages. This ensures a more efficient capacity utilizing. However the modeling of reservation and transmission of BE messages is out of scope of this paper.

3.3 Analytical Model

The message arrival process during each slot is Poisson at each SS. The duration of a transmission slot is τ. We express the number of arrivals in messages per time unit. The mean number of arriving i-messages per time unit is denoted by λ_i. Hence the overall arrival rate is $\lambda = \sum_{i=1}^{N} \lambda_i$. The messages are assumed to be of fixed length. Therefore, b_i denotes the size of an i-message, i.e. the number of packets (transmission slots) in a message arriving to SS i. The arrival processes and the message sizes (in transmission slots) at the different SSs are assumed to be mutually independent.

Denote the duration of the DL and UL sub-frames by T_d and T_u, respectively. T_{ri} stands for the duration of the RI and T_{ud} is the maximum available duration of the uplink data transmission in a frame. It follows that:

$$T_u = T_{ri} + T_{ud}.$$

The transmission time of a BW-Req is α. The reservation interval of each frame consists of K polling slots (transmission opportunities), whose sizes equal to the transmission time of a BW-Req. Hence $T_{ri} = K\alpha$ and we get:

$$T_{ud} = T_u - K\alpha. \tag{1}$$

Since in one frame exactly one i-message can be transmitted on uplink from every SSs, for the optimal capacity allocation of the SSs holds:

$$T_{ud} = \sum_{i=1}^{N} b_i. \tag{2}$$

3.4 Model Assumptions

We denote the *utilization* of SS i by ρ_i. Since each SS gets a chance to transmit on UL at most one message in each frame, we obtain for the utilization of SS i:

$$\rho_i = \lambda_i T_f. \tag{3}$$

Additionally, we formulate the following assumptions of our model:

A.1: The following relation holds for the arrival rate of each SS i:

$$\rho_i = \lambda_i T_f < 1, \quad i = 1, \ldots, N. \tag{4}$$

This relation ensures the stability of the model.

A.2: The time of BS processing including scheduling is negligible.

A.3: The channel propagation time is negligible.

A.4: The transmission channels are error-free.

A.5: BW-Req for i-messages arriving during RI can be sent first time in the RI of the next frame.

4 Overall Delay Analysis

The overall delay of an i-message arises mainly due to waiting of the i-message in the outgoing buffer of SS i to get access for successfully sending bandwidth request (waiting for reservation) and the corresponding queuing in the individual BS buffer of SS i (waiting for scheduling).

4.1 Overall Delay Definition

We define the *overall delay* (W_i) of the tagged i-message as the time interval spent from its arrival into the outgoing buffer of SS i up to the end of its successful transmission in the UL. It is composed of several parts:

$$W_i = W_i^r + \alpha + W_i^s + W_i^t + b_i \tau, \tag{5}$$

where W_i^r is the reservation delay, which is defined as the time interval from the i-message arrival to SS i until the start of successful transmission of the corresponding BW-Req to the BS.

α is the transmission time of a BW-Req.

We define the *scheduling time of the tagged i-message* as the the end of the polling slot, when the tagged i-message is scheduled by BS for transmission on UL in the next frame.

W_i^s is the scheduling delay, which is defined as the time interval from the end of sending a BW-Req of the tagged i-message to its scheduling time.

W_i^t is the transmission delay, which is defined as the time interval from the scheduling time of the tagged i-message to the start of its successful transmission in the UL sub-frame.

$b_i \tau$ is the transmission time of an i-message.

4.2 Reservation and Scheduling Delays

We consider the 2 most important terms of the overall delay (reservation and scheduling delays) together, since it results in a simpler queueing model as treating them separately.

Since SS i has an individual request buffer in BS and a fixed bandwidth for UL transmission in each frame assigned to it, the statistical behavior of a particular SS is independent of the behavior of the other SSs. Therefore the stochastic behavior of a particular SS can be modeled by an individual queueing model.

In the queueing model for the reservation and scheduling delays W_i^t does not need to be taken into account. Hence in this queueing model the service of the tagged i-message starts at its scheduling time, i.e. when the BS schedules that message for transmission on UL in the next frame. In case of empty individual BS buffer of SS i this happens at the end of successful BW-Req transmissions from that SS. Hence in this queueing model the busy periods can start only at the end of successful BW-Req transmissions from SS i. As SS i has fixed bandwidth for UL transmission in each frame assigned to it, the service time is T_f. Thus the appropriate model is an M/D/1 queueing model, in which the service time equals T_f. Furthermore we observe that the service of the arriving i-message can not start until the next successful BW-Req transmissions from SS i even if the individual BS buffer of SS i is empty. Although this is a vacation-like property, we rather apply the approach of [9] by means of the residual service time, since it does not need any higher moments and hence it is simpler.

Applying the mean delay formula of the approach of the residual service time in our model with the corresponding parameters leads to

$$E\left[W_i^r + W_i^s\right] = E\left[W_i^0\right] + \frac{\lambda_i T_f^2}{2(1-\lambda_i T_f)}, \tag{6}$$

where W_i^0 is the *initial message delay*, which is the sum of the reservation and scheduling delays conditioning on the fact that the arriving i-message sees the system empty.

We remark here that (6) is an approximation. The approach of the residual service time – exactly as the vacation model approach with exhaustive service – assumes that the service is work conserving as far as there are i-messages in the system. However if there are i-messages waiting for reservation when the individual BS buffer of SS i becomes empty then the principle of work conserving does not hold any more for this model, because the service stops.

4.3 Initial Message Delay – In General

We assume that in stationary situation the successful BW-Req transmission at SS i in a polling slot has a constant probability, $p_i^{st} > 0$.

In the following we introduce several quantities in order to determine $E\left[W_i^0\right]$ in our model. We define W_i^{rs} as the time interval from the begin of first try of sending a BW-Req of the tagged i-message until the start of successful transmission of the corresponding BW-Req to the BS. Due to the constant probability of the successful BW-Req transmission at SS i in a polling slot W_i^{rs} is geometric in terms of the number of polling slots. More precisely its distribution is given as:

$$P\left\{W_i^{rs} = \lfloor\frac{n}{K}\rfloor T_f + (\frac{n}{K})^* K\alpha\right\} = (1-p_i^{st})^n p_i^{st}, \quad n \geq 0, \tag{7}$$

where $\lfloor c \rfloor$ and $(c)^*$ stand for the integral part and the fractional part of c, respectively.

By definition W_i^{rb} is the interval seen by a first arriving i-message after a successful BW-Req transmission until the begin of first try of sending the BW-Req from that SS. Due to the empty system condition W_i^0 is given as

$$W_i^0 = W_i^{rb} + W_i^{rs}. \tag{8}$$

After a successful BW-Req transmissions until the begin of first try of sending a BW-Req from SS i all arrivals occurs only in the last T_f part. Hence

$$E\left[W_i^{rb}\right] = \frac{T_f}{2}. \tag{9}$$

It is shown in the Appendix that

$$E\left[W_i^{rs}\right] = \frac{\left(1-p_i^{st}\right)^K}{1-\left(1-p_i^{st}\right)^K} T_f + \frac{\left(1-p_i^{st} K \left(1-p_i^{st}\right)^{K-1} - \left(1-p_i^{st}\right)^K\right) \frac{1-p_i^{st}}{p_i^{st}}}{1-\left(1-p_i^{st}\right)^K} \alpha. \tag{10}$$

Applying (9) and (10) in (8) leads to

$$\begin{aligned} E\left[W_i^0\right] = &\frac{T_f}{2} + \frac{\left(1-p_i^{st}\right)^K}{1-\left(1-p_i^{st}\right)^K} T_f \\ &+ \frac{\left(1-p_i^{st} K \left(1-p_i^{st}\right)^{K-1} - \left(1-p_i^{st}\right)^K\right) \frac{1-p_i^{st}}{p_i^{st}}}{1-\left(1-p_i^{st}\right)^K} \alpha. \end{aligned} \tag{11}$$

For high traffic load p_i^{st} can be approximately determined by means of the independent conditional collision probability assumption proposed by Bianchi [11], which leads to a nonlinear equation (see also [13]).

However in other traffic ranges it does not hold, since during the collision resolution process the SSs influences each other. Thus in general the determination of p_i^{st} is a difficult task. Therefore, as a first analytic approximation, we consider a simplified symmetric model to determine $E\left[W_i^0\right]$.

4.4 Initial Message Delay – Symmetric System

Let K per frame equal to 1 and we set the message sizes of all SSs b_i equal to 1 packet. Further we set all λ_i values equal, which makes the system symmetric in terms of the message arrival flows.

The performance of the BEB collision resolution protocol should be optimized for the considered system settings. In [13] after the extensive analysis of the BEB operation it was established that the optimal value of the BEB parameter W

(initial contention window) should be equal to $2N - K$, where N – number of the nrtPS SSs in the system. We remark here that the polling slots of W can be distributed over more frames. The second BEB parameter m (backoff stage) should be equal to 0 for the optimal BEB protocol.

Therefore, the optimized BEB is reduced to the Aloha protocol [13], where each backlogged SS (the one that has at least one message ready for transmission) chooses one of W polling slots following the message arrival uniformly. In case of collision the SS repeats the choice of a random polling slot to retransmit its BW-Req until the transmission is finally successful. Once BW-Req is successfully transmitted to the BS in a polling slot, the queue of messages that belong to the corresponding SS is updated at the BS. Therefore, the information of all the messages accumulated during the contention process is transfered to the BS and the service starts.

In order to establish the initial message delay of a tagged SS i we consider the sequence of service times, i.e. the ends of the polling slots in the frame following the one where a BW-Req was transmitted successfully. We conclude, that as arrival flow is Poisson, the newly arrived message firstly waits for $\frac{T_f}{2}$ time before the first transmission attempt of a BW-Req according to the PASTA property (for a more thorough explanation see [19]). Then the contention process starts, which adds to the initial delay some random number of frames. Below we give an estimation on this random number.

We consider the following simple linear feedback model. Notice that in terms of bandwidth requesting each SS may be either *thinking* or *backlogged.* The thinking SS has no message ready for transmission and generates one during a frame with the probability $z = 1 - e^{-\frac{\lambda}{N}Tf}$. Once a new message is generated, the SS enters the backlogged state where no new arrivals are possible. This corresponds to the real system where after the first arrival an SS starts the contention process and the subsequent arrivals are irrelevant to establish the sought contention delay. Once the transmission of a BW-Req is successful, the SS enters the thinking state and is able to generate new messages. In each polling slot the backlogged SS attempts to transmit a BW-Req with the probability $p = \frac{2}{W+1}$.

We describe the considered linear feedback model [20] with a Markov chain consisting of $N + 1$ states (see Figure 3). Each state corresponds to the number of the backlogged SSs in the system at the moment of the service time $N^{(t)}$. The transition probabilities for the considered Markov chain are as follows:

$$p_{ij} = \Pr\{N^{(t+1)} = j | N^{(t)} = i\} = \qquad (12)$$

$$= \begin{cases} 0,\ j \leq i-2 \\ ip(1-p)^{i-1}(1-z)^{N-i+1},\ j = i-1 \\ ip(1-p)^{i-1}(N-i+1)z(1-z)^{N-i}+ \\ +(1-ip(1-p)^{i-1})(1-z)^{N-i},\ j = i \\ ip(1-p)^{i-1}\begin{pmatrix} N-i+1 \\ j-i+1 \end{pmatrix} z^{j-i+1}(1-z)^{N-j}+ \\ +(1-ip(1-p)^{i-1})\begin{pmatrix} N-i \\ j-i \end{pmatrix} z^{j-i}(1-z)^{N-j},\ j \geq i+1. \end{cases}$$

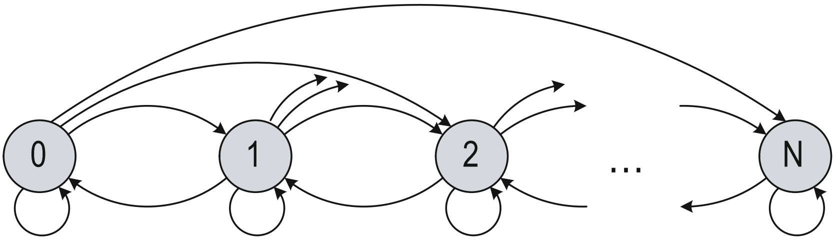

Fig. 3. Markov chain for linear feedback model

It may be shown that the considered chain is finite and irreducible for $p, z > 0$. Therefore, a stationary probability distribution always exists. This distribution may be obtained, for instance, by solving a system of $N + 1$ linear equations:

$$\begin{cases} P_j = \sum\limits_{i=0}^{N} P_i p_{ij}, \ j = 0, 1, \ldots, N \\ \sum\limits_{i=0}^{N} P_i = 1. \end{cases} \tag{13}$$

Using the stationary probability distribution one may obtain the average number of the backlogged SSs

$$B = \sum_{n=1}^{N} n P_n \tag{14}$$

and the stationary success probability

$$S = \sum_{n=0}^{N} s(n,p) P_n, \tag{15}$$

where $s(n,p) = np(1-p)^{n-1}$. Finally, the mean delay in the considered linear feedback model is given by the Little's result, that is $D = \frac{B}{S}$.

Combining the above, the initial message delay in the symmetric system is given by the following expression:

$$E\left[W_i^0\right] = T_f(D + \frac{1}{2}). \tag{16}$$

4.5 Transmission Delay – Symmetric System

Remember, that each SS has a fixed position in the uplink subframe, that is, the transmission delay of SS i is $(i-1)\tau$. Summarizing it over every SSs yields the transmission delay under symmetric settings as

$$W_i^t = \frac{1}{N} \sum_{i=1}^{N} (i-1)\tau = \tau \frac{N-1}{2}. \tag{17}$$

4.6 Mean Overall Message Delay

Applying (5) in symmetric system, the mean overall message delay is given as:

$$E\left[W_i\right] = E[W_i^r + W_i^s] + \alpha + E\left[W_i^t\right] + \tau. \tag{18}$$

Accounting for (18), (6), (16) and (17) the mean overall message delay for the symmetric system can be expressed by:

$$E\left[W\right] = T_f(D + \frac{1}{2}) + \frac{\frac{\lambda}{N}T_f^2}{2(1 - \frac{\lambda T_f}{N})} + \tau\frac{N+1}{2} + \alpha. \tag{19}$$

5 Simulation Results

In order to validate the considered analytical model a simulation program for IEEE 802.16 MAC was developed. The program is a time-driven simulator that accounts for the discussed restrictions on the considered system model. The applied simulation parameters of IEEE 802.16 MAC and PHY, which follows [21], are summarized in Table 1.

Table 1. Basic IEEE 802.16 simulation parameters

Parameter	Value
PHY layer	OFDM
Frame duration (T_f)	5 *ms*
DL/UL ratio	50:50
Channel bandwidth	7 *MHz*
MCS	16 QAM $^3/_4$
Packet length	512 *Byte*
BW-Req duration (α)	0.17 *ms*

For the purposes of simplicity we again restrict our practical explorations to the symmetric system case. This enables a better visibility of the below comparison results.

In Figure 4 we conduct the comparison of the overall delay for the practical system and the established theoretical estimation. We set the number of nrtPS SSs in the system N equal to 6 and run each simulation point for approximately 10 minutes of the real time. Firstly, we notice that the derived analytical expression gives a good estimation on the realistic overall delay value.

Another interesting observation is the intricate shape of the theoretical overall delay curve. Remember, that the analytic formulas are approximations, since the work conserving does not hold any more for this model if there are i-messages waiting for reservation when the individual BS buffer of SS i becomes empty. Therefore the i-messages actually waiting for reservation experience an additional waiting time compared to the work conserving case, where the service

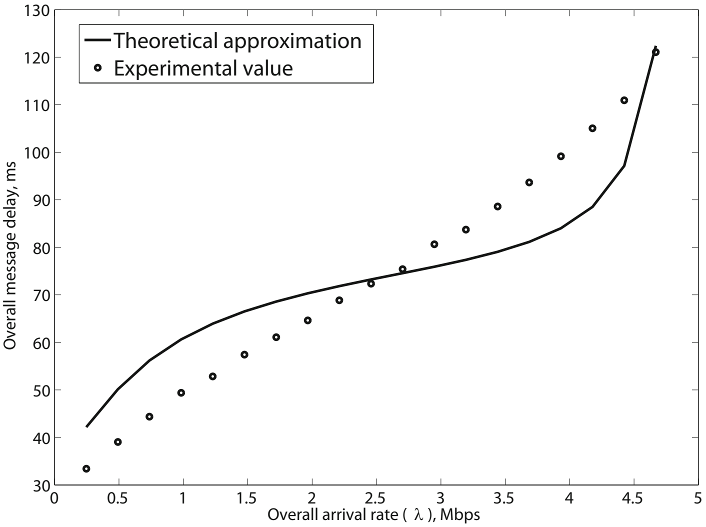

Fig. 4. Verification of the derived model for the symmetric system case

would continue. This supplementary waiting time of the tagged i-message takes until the next successful BW-Req transmission from SS i. It follows that our approach with the work conserving assumptions underestimates the real waiting time. On the other hand, the analytical approach overestimates the reservation time since the first message in a batch transmitted in one BW-Req waits longer than the others. Therefore we see that the shape of the curve changes. However the result gives a good approximation for the practical value.

6 Conclusion

In this paper we have developed an analytical approximation for the mean overall message delay of nrtPS traffic in IEEE 802.16 wireless network. This model accounts for both reservation delay and scheduling delay components and enables asymmetric Poisson arrival flows and different message sizes. More importantly, the contention-based broadcast polling of subscriber stations is the studied bandwidth reservation mechanism.

For symmetric system a simple linear feedback model is used to estimate the contention delay. As our experiments show, the established theoretical overall delay gives a good approximation for the practical values obtained with simulation.

The analytical approach used for symmetric system (the linear feedback model) can be extended for the asymmetric system as well, where arrival flows and/or message sizes are not equal. Moreover, the model may be also extended to account the case of $K > 1$ and the various BEB parameters.

Acknowledgement

Dr. Alexey Vinel acknowledges the support of Alexander von Humboldt Foundation for the funding of this work, which was done in part during his stay in Germany.

References

1. IEEE Std 802.16e-2005, Piscataway, NJ, USA (December 2005)
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. Boxma, O.J., Groenendijk, W.M.: Waiting times in discrete-time cyclic-service systems. IEEE Trans. on Comm. 36(2), 164–170 (1988)
4. 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), 219–233 (1987)
5. Goodman, J., Greenberg, A., Madras, N., March, P.: Stability of binary exponential backoff. Journal of the ACM 35(3), 579–602 (1988)
6. Chlebus, B.: Randomized Communication in Radio Networks. In: Pardalos, P., Rajasekaran, S., Reif, J., Rolim, J. (eds.) Handbook of Randomized Computing, vol. 1, pp. 401–456 (2001)
7. Tsybakov, B.: Survey of ussr contributions to random multiple-access communications. IEEE Transactions on Information Theory 31(2), 143–165 (1985)
8. Tsybakov, B., Mikhailov, V.: Free synchronous packet access in a broadcast channel with feedback. Problems of Information Transmission 14(4), 259–280 (1978)
9. Bertsekas, D., Gallager, R.: Data Networks. Prentice-Hall, Englewood Cliffs (1992)
10. Song, N., Kwak, B., Miller, L.: On the stability of exponential backoff. Journal Research of NIST 108, 289–297 (2003)
11. Bianchi, G.: Performance analysis of the ieee 802.11 distributed coordination function. IEEE Journal on Selected Areas in Communications 18(3), 535–547 (2000)
12. Lin, L., Jia, W., Lu, W.: Performance analysis of ieee 802.16 multicast and broadcast polling based bandwidth request. In: IEEE Wireless Communications and Networking Conference, vol. 1, pp. 1854–1859 (2007)
13. Andreev, S., Turlikov, A., Vinel, A.: Contention-based polling efficiency in broadband wireless networks. In: International Conference on Analytical and Stochastic Modelling Techniques and Applications, vol. 1, pp. 295–309 (2008)
14. Alanen, O.: Multicast polling and efficient voip connections in ieee 802.16 networks. In: 10th ACM Symposium on Modeling, analysis, and simulation of wireless and mobile systems, vol. 1, pp. 289–295 (2007)
15. Paschos, G.S., Papapanagiotou, I., Argyropoulos, C.G., Kotsopoulos, S.A.: A heuristic strategy for ieee 802.16 wimax scheduler for quality of service. In: 45th Congress FITCE (2006)
16. de Moraes, L.F.M., Maciel, P.D.: A variable priorities mac protocol for broadband wireless access with improved channel utilization among stations. In: Int. Telecomm. Symp., vol. 1, pp. 398–403 (2006)
17. Chang, Y.-J., Chien, F.-T., Kuo, C.-C.J.: Delay analysis and comparison of ofdm-tdma and ofdma under ieee 802.16 qos framework. In: IEEE Global Telecomm. Conf (GLOBECOM), vol. 1, pp. 1–6 (2006)

18. Iyengar, R., Iyer, P., Sikdar, B.: Delay analysis of 802.16 based last mile wireless networks. IEEE Global Telecommunications Conference 5, 3117–3127 (2005)
19. Saffer, Z.s., Andreev, S.: Delay analysis of ieee 802.16 wireless metropolitan area network. In: Int. Workshop on Multiple Access Communications (MACOM) (2008)
20. Kleinrock, L.: Queueing Systems: Volume II – Computer Applications. Wiley Interscience, Hoboken (1976)
21. Sivchenko, D., Bayer, N., Xu, B., Rakocevic, V., Habermann, J.: Internet traffic performance in ieee 802.16 networks. In: European Wireless (2006)

A Mean of W_i^{rs}

Using (10) the mean of W_i^{rs} can be expressed for $0 < p_i^{st} < 1$ as

$$E\left[W_i^{rs}\right] = \sum_{j=0}^{\infty} \left(jT_f \sum_{n=jK}^{(j+1)K-1} \left(1-p_i^{st}\right)^n p_i^{st} + \sum_{n=jK}^{(j+1)K-1} (n-jK)\,\alpha \left(1-p_i^{st}\right)^n p_i^{st} \right). \quad (20)$$

Rearranging results in

$$\begin{aligned}
E\left[W_i^{rs}\right] &= \sum_{j=0}^{\infty} \Bigg(jT_f \left(1-p_i^{st}\right)^{jK} \sum_{n-jK=0}^{K-1} \left(1-p_i^{st}\right)^{n-jK} p_i^{st} \\
&\quad + \alpha p_i^{st} \left(1-p_i^{st}\right)^{jK} \sum_{n-jK=0}^{K-1} (n-jK) \left(1-p_i^{st}\right)^{n-jK} \Bigg) \qquad (21) \\
&= \sum_{j=0}^{\infty} \Bigg(jT_f \left(1-p_i^{st}\right)^{jK} \left(1-\left(1-p_i^{st}\right)^K\right) \\
&\quad + \alpha \left(1-p_i^{st}K\left(1-p_i^{st}\right)^{K-1} - \left(1-p_i^{st}\right)^K\right) \frac{1-p_i^{st}}{p_i^{st}} \left(1-p_i^{st}\right)^{jK} \Bigg) \\
&= T_f \left(1-\left(1-p_i^{st}\right)^K\right) \sum_{j=0}^{\infty} j \left(\left(1-p_i^{st}\right)^K\right)^j \\
&\quad + \alpha \left(1-p_i^{st}K\left(1-p_i^{st}\right)^{K-1} - \left(1-p_i^{st}\right)^K\right) \frac{1-p_i^{st}}{p_i^{st}} \sum_{j=0}^{\infty} \left(\left(1-p_i^{st}\right)^K\right)^j \\
&= T_f \left(1-\left(1-p_i^{st}\right)^K\right) \frac{\left(1-p_i^{st}\right)^K}{\left(1-\left(1-p_i^{st}\right)^K\right)^2} \\
&\quad + \alpha \left(1-p_i^{st}K\left(1-p_i^{st}\right)^{K-1} - \left(1-p_i^{st}\right)^K\right) \frac{1-p_i^{st}}{p_i^{st}} \frac{1}{1-\left(1-p_i^{st}\right)^K} \\
&= \frac{\left(1-p_i^{st}\right)^K}{1-\left(1-p_i^{st}\right)^K} T_f + \frac{\left(1-p_i^{st}K\left(1-p_i^{st}\right)^{K-1} - \left(1-p_i^{st}\right)^K\right) \frac{1-p_i^{st}}{p_i^{st}}}{1-\left(1-p_i^{st}\right)^K} \alpha .
\end{aligned}$$

We remark here that $p_i^{st} = 1$ implies $W_i^{rs} = 0$ and thus (21) holds also for $p_i^{st} = 1$.