Contention-Based Polling Efficiency in Broadband Wireless Networks

Sergey D. Andreev[1], Andrey M. Turlikov[1], and Alexey V. Vinel[2]

[1] St. Petersburg State University of Aerospace Instrumentation,
Bolshaya Morskaya street, 67,
190000, St. Petersburg, Russia
{corion,turlikov}@vu.spb.ru
http://suai.ru/

[2] St. Petersburg State University of Information Technologies, Mechanics and Optics,
Vasilievsky Island, Birjevaja Linija, 4,
199034, St. Petersburg, Russia
vinel@ieee.org
http://www.ifmo.ru/eng/

Abstract. This paper addresses the performance of the contention-based polling techniques at the bandwidth reservation stage of IEEE 802.16 standard. A general proposition is proved, which establishes that the grouping of users in the random multiple access system does not change its capacity. Broadcast and multicast polling mechanisms are then considered, for which the throughput and the rate of the truncated binary exponential backoff algorithm are calculated for the lossy and the lossless system types, respectively. It is shown, that subject to proper optimization the performance of the aforementioned algorithm is the same for both system types. The efficiency of the symmetric user grouping is finally studied, which demonstrates that a negligible performance gain may be achieved for the cost of the increased IEEE 802.16 overhead.

1 Introduction and Background

IEEE 802.16 standard [1] defines a high-speed access system supporting multimedia services. 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. IEEE 802.16 is likely to emerge as an outstanding cost-competitive technology mainly for its longer range and sophisticated *quality-of-service* (QoS) support at the MAC layer.

Many research papers concentrate on the performance evaluation of the various IEEE 802.16 features. In particular, the bandwidth requests transmission by a system user to reserve a portion of the channel resources is frequently addressed. A detailed description of the reservation techniques is known from the fundamental work in [2]. The standard allows a *random multiple access* (RMA) scheme at the reservation stage and implements the truncated *binary exponential backoff* (BEB) algorithm for the purposes of the collision resolution.

K. Al-Begain, A. Heindl, and M. Telek (Eds.): ASMTA 2008, LNCS 5055, pp. 295–309, 2008.

The asymptotic behavior of the BEB algorithm was substantially addressed in the literature. In [3] it was shown that the BEB algorithm is *unstable* in the infinitely-many users case. By contrast, [4] shows that the BEB is *stable* for any finite number of users, even if it is extremely large, and sufficiently low input rate. These seemingly controversial results demonstrate the two alternative approaches to the analysis of an RMA algorithm [5]. The former is the *infinite population* model, which studies the ultimate performance characteristics of an RMA algorithm. The latter is the *finite population* model that addresses the limits of the practical algorithm operation. An exhaustive description of both models may be found in [6] and [7].

Both finite and infinite models require a framework of additional assumptions, which makes the analysis mathematically tractable. The set of assumptions given by [8], [9] and in Section 3 has nowadays become classical and evolved into a *reference* RMA model. The performance of the BEB algorithm in the framework of the reference model is addressed in [10], which allowed a deeper insight into its operation. In [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. In Section 2 we give a brief description of IEEE 802.16 features. It is known that the BEB algorithm may be adopted for both *broadcast* and *multicast* user polling. In case of multicast polling the set of all system users is divided into smaller subsets. The efficiency of broadcast and multicast polling was extensively studied in [12]. Some practical aspects of the BEB application for the delay-sensitive traffic were considered in [13].

The motivation behind this paper is to show that despite the fact that for some scenarios multicast polling results in a slightly better system performance, like it is claimed in [12] and [13], the gain is practically negligible when all the users share the similar QoS requirements. In order to verify this hypothesis, we firstly address the infinite population model in Section 4 and show that the RMA *capacity* (see [14] and [15]), cannot be increased by the grouping of users.

In Section 5 we study the BEB algorithm performance in a practical finite population model. Further, by using various analytical techniques we mathematically express the possible gain from the use of broadcast/multicast polling for the different types of the system. The Conclusion summarizes the paper.

2 Standard Overview

IEEE 802.16 standard specifies PHY and MAC layers and supports two modes of operation: the mandatory point-to-multipoint mode (PMP) and the optional mesh mode. The MAC layer is subdivided into three hierarchical sub-layers. Through the *convergence sub-layer* IP, ATM and Ethernet traffic types are supported. Five levels of QoS are specified within *MAC sub-layer*, which correspond to the QoS classes. MAC data packets can be variable-length with concatenation and fragmentation mechanisms supported. *Privacy sub-layer* performs the encryption of the data packets together with the other cryptographic functions.

The basic IEEE 802.16 architecture assumes that there are one *base station* (BS) and one or more subscriber stations, which are referred to as *users* in what follows. The packet exchange between the BS and the users is assumed to be via separate channels. A *downlink* channel is from the BS to the users and the *uplink* channel is in the reverse direction. Therefore, there is no particular connection associated with the downlink channel, while in the uplink channel all the connections from all the users are multiplexed.

IEEE 802.16 defines two duplexing mechanisms the channels: time division duplex (TDD) and frequency division duplex (FDD). In the TDD mode the frame is separated into the downlink and the uplink parts. The simplified structure of the MAC frame in the TDD mode is shown in Fig. 1. In the FDD mode the users transmit in different sub-bands and do not interfere with each other.

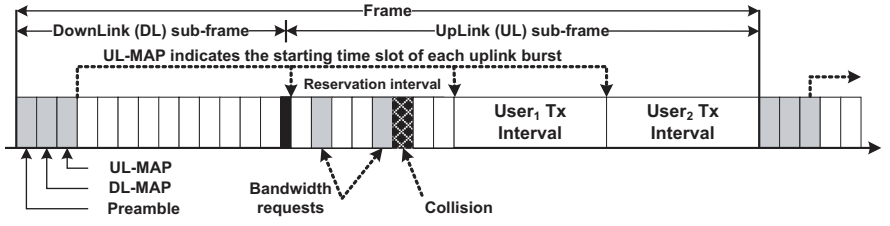

Fig. 1. IEEE 802.16 frame structure in the TDD mode

In the downlink channel the BS is the only station to broadcast packets to all the users. Together with the data packets, the BS also transmits service information about the schedule for each of the users in the uplink channel. This information is incorporated in the UL-MAP message and is used by the users for scheduling their data packets in the uplink channel. To allow the feedback from the users the BS also specifies a portion of channel resources as the *reservation interval*. During this interval the users transmit their *bandwidth requests* (or requests, for short), which are then processed by the BS.

The access procedure of the users to the reservation interval could be either *contention-based* or *contention-free*. The latter is referred to as unicast polling, where the BS assigns a transmission opportunity (which is referred to as *slot* below) to each user for its bandwidth requests. The former comprises two mechanisms, namely, multicast and broadcast polling. During broadcast polling the users send their bandwidth requests by choosing one of all the available slots. In case of multicast polling the users are polled in groups and within a group the rules of broadcast polling apply. During the contention-based access the request collisions may occur, which are resolved following the truncated binary exponential backoff algorithm. The *piggybacking* mode allows a user to attach further bandwidth requests to its data packets, once bandwidth for the first packet has been granted by the BS.

3 Reference System Model

In this section the reference RMA model [8], [9] is discussed that simplifies the below derivations. The system time is divided into adjacent frames of the equal duration. The frames are enumerated with integer and nonnegative numbers. Suppose there are M users in the system. We formulate additional assumptions about the way the requests arrive into the system and are transmitted.

Assumption 1. According to IEEE 802.16 standard each user may potentially establish multiple connections with the BS using different negotiated QoS parameters, and a bandwidth request can be issued on a per-connection or a per-station basis. In what follows we assume that each user has only one connection at a time and all the connections belong to the same QoS class.

Assumption 2. Each frame comprises K equal contention slots for the request transmissions. K is constant throughout the system operation.

Assumption 3. In each slot one of the following situations may occur:
– exactly one user transmits its request (*success*);
– none of the users transmit the request (*empty*);
– two or more users transmit their requests simultaneously, which results in the corruption of all the requests at the BS (*collision*).

Assumption 4. The uplink channel is noise-free. Therefore, the BS faultlessly determines, which situation occurred in a slot. If only one user transmits, then the BS always decodes the bandwidth request successfully.

Assumption 5. No piggybacking is used and for each arrived data packet a separate bandwidth request is generated. As we concentrate on the bandwidth reservation process, we assume the virtual input flow of requests into the system.

Assumption 6. By monitoring user activity in the frame $t-1$ the BS makes a schedule for the uplink sub-frame of the frame t and broadcasts this schedule in the downlink sub-frame of the frame t. A user receives the feedback from the request transmission in the frame $t-1$ by the beginning of the frame t.

According to the standard this is not the case. Feedback information is not explicitly transmitted to a user. A special request timeout is used to wait for the uplink grant from the BS, and only if it is expired, the request transmission is considered corrupted. We make this 'immediate' feedback assumption for the simplicity of the analysis only. All the forthcoming derivations may be generalized for the case of the 'delayed' feedback.

Assumption 7. The downlink channel is noise-free. Therefore, a user faultlessly receives the schedule and the request transmission feedback from the BS.

Assumption 8. Denote the random number of the new request arrivals to the user i in the frame t by $X_i^{(t)}$. For all $t \geq 0$ and $i = 1, \ldots, M$ the random variables $X_i^{(t)}$ are independent and identically distributed (i.i.d.). Assume also that at

most one new request arrives to a user per frame with the probability y. Thus, $E[X_i^{(t)}] = y$ for all $t \geq 0$ and $i = 1, \ldots, M$, as well as $E[\sum_{i=1}^{M} X_i^{(t)}] = My \triangleq \Lambda$. The value of Λ is hereinafter referred to as the cumulative *input rate* and the considered input flow constitutes a Bernoulli flow.

4 Infinite User Population

Following the approach from [5] we allow the number of users in the system M to increase infinitely and the probability of a request arrival y to decrease simultaneously so that their product remains constant, that is $My = const = \Lambda$. Then the limit of the cumulative arrival process given by Assumption 8 is Poisson, i.e. $\lim_{M\to\infty} \Pr\{\sum_{j=1}^{M} X_j^{(t)} = i\} = \frac{\Lambda^i}{i!} e^{-\Lambda}$. Below we make the basic definitions and introduce *lossy* and *lossless* system types as follows.

4.1 Lossy System

Definition 1. The *RMA algorithm* A from the class of algorithms for the lossy system $A \in \mathcal{A}_{lossy}$ is defined as a rule that allows a user with a pending request to determine whether it should transmit this request in the following slot s or *discard* it. If a request is discarded then the corresponding data packet is lost [16].

Definition 2. We introduce a random variable $Z^{(t)}$, which is the number of the successful request transmissions in a frame comprising K slots. Clearly, $Z^{(t)} \in \{0, 1, \ldots, K\}$. Define the random variable $\Psi_A(K, \Lambda, s) \triangleq \frac{\sum_{j=0}^{s} Z^{(t)}}{sK}$. The limit of this expression for s, if it exists, represents the *output rate* per slot of the algorithm A in the lossy system, that is $\Psi_A(K, \Lambda) \triangleq \lim_{s\to\infty} \Psi_A(K, \Lambda, s)$.

Definition 3. The *throughput* of the algorithm A in the lossy system is the maximum achievable output rate for all the input rates, which implies:

$$T_A(K) \triangleq \sup_{\Lambda} \Psi_A(K, \Lambda). \tag{1}$$

Definition 4. The *capacity* of the lossy system is the maximum throughput over the class $\mathcal{A}_{lossy}(K)$ of the RMA algorithms with K slots per frame:

$$\mathcal{C}_{lossy}(K) \triangleq \sup_{A\in\mathcal{A}_{lossy}(K)} T_A(K)\,. \tag{2}$$

Notice, that the throughput value characterizes the behavior of an RMA algorithm, whereas the capacity gives the ultimate performance threshold for the entire lossy system.

4.2 Lossless System

Definition 5. The *RMA algorithm* A from the class of algorithms for the lossless system $A \in \mathcal{A}_{lossless}$ is defined as a rule that allows a user with a pending request to determine whether it should transmit this request in the following slot s. Notice, that no discard rule is specified and, consequently, requests are never lost.

Definition 6. The *request delay* for an RMA algorithm is the time interval from the moment of the request generation to the moment of its successful transmission. The delay $\delta_A(K, \Lambda)$ is a random variable. We inject a new request into the system at the randomly chosen slot s, and denote the delay of this request as $\delta_A^{(s)}(K, \Lambda)$.

Definition 7. The *mean delay* (referred to as virtual mean delay in [7]) is defined as:

$$D_A(K, \Lambda) \triangleq \overline{\lim_{s\to\infty}} E[\delta_A^{(s)}(K, \Lambda)]. \tag{3}$$

Definition 8. The transmission *rate* (tenacity) of the algorithm A in the lossless system is the maximum input rate that can be sustained by the algorithm with finite request delay:

$$R_A(K) \triangleq \sup_{\Lambda}\{\Lambda : D_A(K, \Lambda) < \infty\}. \tag{4}$$

Definition 9. The *capacity* of the lossless system is the maximum possible rate over the class $\mathcal{A}_{lossless}(K)$ of the RMA algorithms with K slots per frame:

$$\mathcal{C}_{lossless}(K) \triangleq \sup_{A\in A_{lossless}(K)} R_A(K) \ . \tag{5}$$

The exact value of the capacity is not yet established. However, the best known upper bound on the capacity $\mathcal{C}_{lossless}(1)$ was established in [14] and is shown to be $\overline{\mathcal{C}}_{lossless}(1) = 0.587$. The best known part-and-try RMA algorithm was proposed in [17] and its rate is $R_{pt} = 0.487$. In subsequent years it was slightly improved, but the core idea of the algorithm remained unchanged.

Notice again, that the rate value characterizes the behavior of an RMA algorithm, whereas the capacity gives the ultimate performance threshold for the entire lossless system.

4.3 User Grouping Analysis

Here we concentrate on showing that the grouping of users does not increase the ultimate measure of the system performance, namely, its capacity. The below arguments may be repeated similarly for both lossy and lossless types of the system. Below we demonstrate the proof for the lossless system, but omit the lower 'lossless' index at $\mathcal{A}$ and $\mathcal{C}$ as redundant.

1. Firstly, consider the RMA system without the framing structure. The system time is divided into equal slots and a user is restricted to start its request transmission in the beginning of a slot. The RMA algorithm A in this system $\mathcal{A}_{slotted}$ may again be defined as a rule that allows a user with a pending request to determine whether it should transmit this request in the following slot s. The feedback of the user transmission is available by the beginning of the next slot $s+1$.
2. Now we additionally divide the system time into frames with each frame comprising some integer and constant number of slots K. However, the feedback is still available after each slot. It is assumed that all the system users monitor the system activity from the start of its operation. Therefore, all the users determine the situation in each slot similarly and the introduction of frames neither improves nor degrades the system performance. The conclusion we draw from this fact is that the set of all the RMA algorithms for this system $\mathcal{A}_{framed}$ coincides with the set of algorithms for the slotted system, that is $\mathcal{A}_{framed} = \mathcal{A}_{slotted} \triangleq \mathcal{A}(1)$. Analogously to the Definition 9 we define the capacity of the framed system as $\mathcal{C}(1) \triangleq \sup_{A \in A(1)} R_A(1)$.
3. We change the feedback availability for the framed system and let a user know the consequences of a request transmission only in the beginning of the next frame, that is, once in K slots. An alternative system with 'delayed' feedback was considered in [18]. We define the RMA algorithm A for this system $\mathcal{A}(K)$ as before and conclude that with the restriction on the feedback availability the set of all possible RMA algorithms is narrowed in comparison to the respective set for the framed system, which yields $\mathcal{A}(K) \subset \mathcal{A}(1)$.

   From the above and the two definitions of capacity $\mathcal{C}(1)$ and $\mathcal{C}(K)$ (5) it immediately follows that $\mathcal{C}(K) \leq \mathcal{C}(1)$.
4. To any algorithm A from the set $\mathcal{A}(1)$ an algorithm A^* may be put into correspondence that belongs to $\mathcal{A}(K)$, such as $R_{A^*} = R_A$. For this it is sufficient to split all the users of the framed system into K equal groups and restrict the slots available for each group to one slot per frame. For instance, group number one monitors and transmits in the first slot of each frame, group number two - in the second, etc. Therefore, for each group the feedback is available at the beginning of the next slot, dedicated to this particular group, which corresponds to the slotted system.
5. From the definition of the capacity and the above (see 3, 4) it follows that $\mathcal{C}(1) = \mathcal{C}(K)$, that is, the capacity does not change for the framed system. Moreover, when all the system users are already split into equal groups with L slots for each of them, the capacity does not change either, i.e. $\mathcal{C}(1) = \mathcal{C}(L) = \mathcal{C}(K)$. We conclude that the grouping of users leaves the system capacity unchanged.

5 Finite User Population

In this section we address the contention-based polling performance in the framework of the model from Section 3 for a practical case of the finite number of users M. We narrow the set of all the RMA algorithms to one algorithm which

is specified by IEEE 802.16 standard. This algorithm is the truncated binary exponential backoff (BEB), which steps may be summarized as follows.

5.1 Truncated Binary Exponential Backoff Algorithm

Rule 1.1. If a new bandwidth request arrives to a user in the frame $t-1$ and this user has no other pending requests, it transmits the request in the frame t (transmission attempt). The slot for the request transmission is sampled uniformly from the number of contention slots dedicated to the group the user belongs to. Notice, that in case of broadcast polling the user may choose between all the contention slots K of the frame t, whereas in case of multicast polling the choice is narrowed to L slots of the respective multicast group.

Rule 1.2. If a request is ready for *retransmission* at the beginning of the frame t at its i-th retransmission attempt ($i > 0$), a user chooses a number (*backoff counter*) in the range $\{0, 1, \ldots, 2^{min(m,i)}W - 1\}$ uniformly, where W and m are the parameters of the BEB algorithm, named *initial contention window* and *maximum backoff stage* respectively and i is the number of collisions this request suffered from so far. The user then defers the request retransmission for the chosen number of slots, accounting only for the slots dedicated to its group.

Rule 2.1. If, after receiving the feedback from the BS, the user determines that its last request collided, it increments the collision counter i for this request. If this counter coincides with the maximum allowable number of retransmission attempts Q, then the request together with the corresponding data packet is discarded and the collision counter is reset to $i = 0$.

Rule 2.2. If, after receiving the feedback from the BS, the user determines that the (re)transmission of the last request was successful, it resets the collision counter to $i = 0$.

5.2 Lossy System

We are interested in the derivation of the BEB algorithm throughput T_{BEB} for the case of minimum possible delay. The motivation for this is the performance evaluation of the delay-critical applications (like VoIP in [13]). In order to minimize the delay for both broadcast and multicast polling the the maximum number of retransmission attempts is set to its minimum value, that is $Q = 0$. Therefore, the corresponding throughput value is denoted as T^1_{BEB}, where 1 stands for the single transmission attempt.

Remember, that according to the Bernoulli input flow (see Assumption 8) the value of y represents the probability of a request arrival to a user in a frame. The standard does not define any relationship between the parameters W, m and K. Notice, for example, that if $W < L$ for multicast polling, then some slots are never used during the first retransmission attempt. For this reason, we set $W = \frac{lK}{G} = lL$, where l is a natural number ($l \geq 1$), in order to distribute the retransmission attempts over the available slots for each multicast group uniformly. In case of no retransmission attempts, $l = 1$ and $m = 0$.

Below we address the throughput T^1_{BEB} per slot, which is achievable by the transmission in the contention slots for both broadcast ($G = 1$) and multicast ($G > 1$) polling. Following the approach from [19] we establish the following:

$$T^1_{BEB}(G, y) = \frac{G}{K} \sum_{k=0}^{N} \binom{N}{k} y^k (1-y)^{N-k} \sum_{i=0}^{\min(k,L)} iP(i, k, L), \tag{6}$$

where $P(r, k, L)$ is the probability that r stations out of k active (with at least one pending request) successfully transmit in a frame that comprises K slots. Denote by $F(r, k, L)$ the total number of ways to put k balls into L boxes, conditioning on the fact that exactly r boxes contain one ball. This number may be computed recursively by the following expressions:

$$F(0, 0, L) = 1, F(0, k, 0) = 0, F(0, k, L) = L^k - \sum_{i=1}^{\min(i,L)} F(i, k, L), k > 0,$$

$$F(r, k, L) = \binom{k}{r} \binom{L}{r} r! F(0, k-r, L-r), 0 < r \leq \min(k, L). \tag{7}$$

Thus, the conditional probability $P(r, k, L)$ equals to:

$$P(r, k, L) = \frac{F(r, k, L)}{L^k}. \tag{8}$$

Fig. 2 demonstrates the function T^1_{BEB} for different number of groups G, $K = 8$ and $M = 40$. We observe that multicast polling outperforms broadcast polling for small input rates y, whereas the situation reverses for moderate and high input rates. We also notice that the gap between the cases with $G = 1$ and $G = 8$ is the most significant and shows the maximum possible gain/loss from the use of either of polling techniques. We plot the dependence of this maximum gain/loss on the input rate in Fig. 3.

We conclude that despite the fact that the use of multicast or broadcast polling demonstrates a throughput trade-off for different values of request arrival rate, the maximum possible gain/loss is negligible in comparison to the achievable throughput. Therefore, it is not reasonable to split users into multicast groups for the considered minimum delay case ($Q = 0$), as the gain is minor, but IEEE 802.16 overhead increases as the number of groups grows [1].

The result given by (6) may alternatively be obtained with the following approach, one similar to which was first addressed in [20]. In each slot at most one request may be transmitted. We introduce a random variable $Z^{(i)}$ that is equal to 1 in case of *success* in the slot i and is equal to 0 otherwise. Notice, that as the number of users in each group is constant and the users are independent, it is sufficient to obtain the expectation of the sum $Z^{(i)}$ over L for one group only. Clearly, this expected value gives the sought throughput T^1_{BEB}, that is:

$$T^1_{BEB} = \frac{E[\sum_{i=1}^{L} Z^{(i)}]}{L} = E[Z^{(i)}]. \tag{9}$$

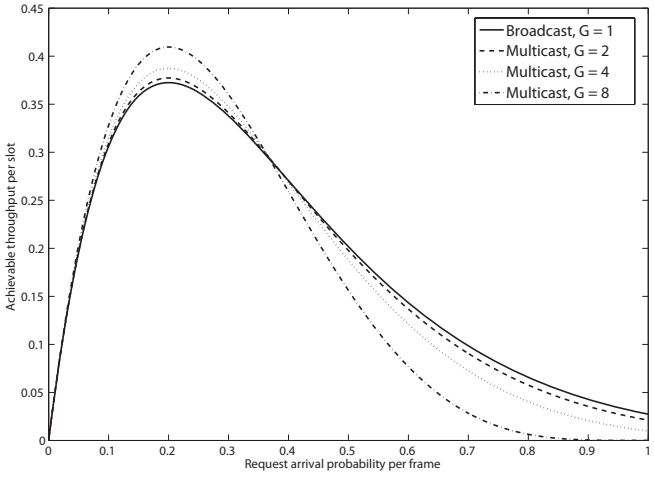

Fig. 2. Contention-based polling efficiency in case of no retransmissions

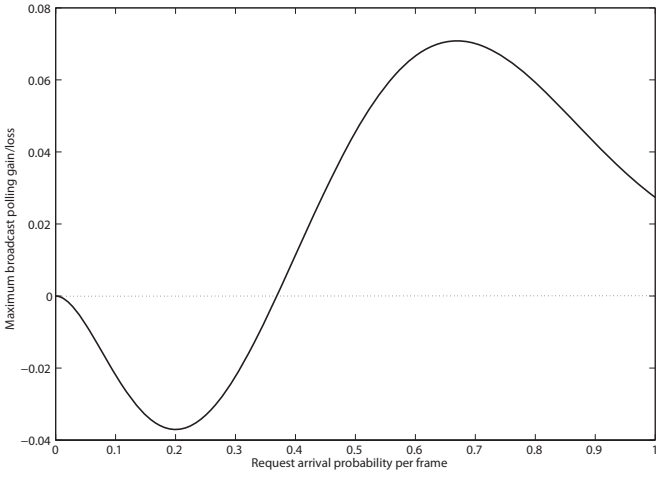

Fig. 3. Maximum gain/loss of the broadcast polling

The expected value of $Z^{(i)}$ represents the probability of the *success* in a slot, which happens *iff* one of N users in a group choses this slot for the request transmission, yielding:

$$T^1_{BEB} = E[Z^{(i)}] = Pr\{Z^{(i)} = 1\} = \frac{yN}{L}(1 - \frac{y}{L})^{N-1}. \tag{10}$$

We notice that the above closed-form expression gives exactly the same result as (6). Additionally, by calculating the first derivative of (10) for y and imposing it equal to 0, we establish the 'optimal' value of the input rate y that results in the maximum throughput value as:

$$y_0 = \frac{L}{N}. \tag{11}$$

The demonstrated approach allows the derivation of a closed-form expression for the maximum broadcast polling gain/loss function (depicted in Fig. 3) as follows:

$$f(y) = \frac{yM}{K}(1 - \frac{y}{K})^{M-1} - \frac{yN}{L}(1 - \frac{y}{L})^{N-1}. \tag{12}$$

5.3 Lossless System

We continue the performance evaluation of the system with finite number of users M under the assumptions given in Section 3. To get a deeper insight into the limitations of the BEB algorithm operation we set the maximum number of request retransmissions Q infinite. This way a request is never discarded and no data packet losses are possible in the system. We are interested in obtaining the rate of the BEB algorithm in the finite population lossless system R_{BEB}.

We introduce the stochastic process $c(s)$ that represents the value of the randomly sampled backoff counter at time s given that the number of collisions suffered by a request so far is $b(s)$. A discrete and integer time scale is also adopted, where s and $s+1$ correspond to the start times of two successive slots. We demonstrate our approach for broadcast polling as the example. All the below derivations may be generalized for the case of multicast polling with N users per group.

We notice, that according to the BEB rules described in Section 5.1 a user after its (re)transmission attempt does not start the backoff process immediately, but rather waits for the beginning of the next frame. Assume, that the (re)transmission attempt occurs in slot s in the frame that consists of K slots. Therefore, the user waits $K - s$ slots before resuming the backoff procedure. At its every retransmission attempt a user may be regarded as choosing the frame to retransmit in first and then choosing one of K slots in this frame. Thus, the number of slots before the (re)transmission in a frame is sampled uniformly in the range $[0, \ldots, K-1]$. Denote the *waiting time counter* as $a(s)$, which accounts for the slots after the (re)transmission attempt by a user and before the start of the next frame.

The considered stochastic process represents a Markov chain analogous to one described in [11] and [21], but with the addition of $K-1$ idle states, which correspond to the possible waiting time counter values. The transition probabilities for these additional states may be computed as follows:

$$\begin{aligned} \Pr\{a(s+1) = k-1 \,|a(s) = k\} &= 1, k = 1, \ldots, K-1, \\ \Pr\{a(s+1) = k \,|b(s) = 0\} &= \frac{1}{K}, k = 1, \ldots, K-1. \end{aligned} \tag{13}$$

Let $b_{i,j} = \lim_{s\to\infty} \Pr\{b(s) = j, c(s) = i\}$, $a_k = \lim_{s\to\infty} \Pr\{a(s) = k\}$, where $i = \{0, \ldots, m\}$, $j = \{0, \ldots, 2^i W - 1\}$ and $k = 1, \ldots, K-1$ is the stationary

distribution of the considered Markov chain. As the probability of a (re) transmission attempt in a slot is equal to $\sum_{i=0}^{m} b_{i,0}$, we establish:

$$a_k = \frac{k}{K-k}\sum_{i=0}^{m} b_{i,0} \Rightarrow \sum_{k=1}^{K-1} a_k = \frac{K-1}{2}\sum_{i=0}^{m} b_{i,0} = \frac{K-1}{2}\cdot\frac{b_{0,0}}{1-p_c}, \tag{14}$$

where p_c is the conditional collision probability, which is equal to the probability that at least one of the remaining $M-1$ users (re)transmits:

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

Accounting for the normalization condition:

$$1 = \sum_{i=0}^{m}\sum_{j=1}^{2^i W} b_{i,j} + \sum_{k=1}^{K} a_k, \tag{16}$$

we notice that the first term is given in [11]. Summarizing, the probability p_t that a user (re)transmits in a randomly chosen slot is readily obtained as:

$$p_t = \sum_{i=0}^{m} b_{i,0} = \frac{2\,(1-2p_c)}{(1-2p_c)\,(W+K) + p_c W\,(1-(2p_c)^m)}. \tag{17}$$

Equations (15) and (17) represent a nonlinear system with two unknowns p_c and p_t, which may be solved numerically. The resulting R_{BEB} value is finally given by the probability of one (re)transmission in a slot:

$$R_{BEB} = Mp_t(1-p_t)^{M-1}. \tag{18}$$

The above approach allows the derivation of the optimal (re)transmission probability value that gives the maximum BEB algorithm rate over all possible pairs of (W, m). It may be shown that this maximum value is reached for $m = 0$. Below we consider the optimal system in more detail.

Substituting $m = 0$ into (17) we obtain that $p_t = \frac{2}{W_0+K}$, where W_0 is the optimal initial contention window value. Notice that (18) closely resembles the expression (10), which is maximized for $\frac{yN}{L} = 1$. Therefore, the expression (18) itself is maximized for $Mp_t = \frac{2M}{W_0+K} = 1$. Finally, W_0 is obtained as $2M-K$, or, accounting for the possible grouping of users:

$$W_0 = 2N - L. \tag{19}$$

It should be emphasized, that the rate of the optimized BEB algorithm with $m = 0$ and W_0 gives precisely the same value as calculated by (10) for the lossy system. However, the usage of the optimal initial contention window W_0 in IEEE 802.16 standard is not straightforward, as it may not be an integer power of two. For this reason we depict the BEB rate for various values of m and different initial contention windows in Fig. 4. We see, that for the example system with $M = 40$, $K = 8$ and broadcast polling, $W_0 = 72$. The BEB rate given by $W = 32$ and $m = 2$ is almost as high as the optimal one. Summarizing, our approach allows the optimization of BEB parameters in terms of the highest achievable rate.

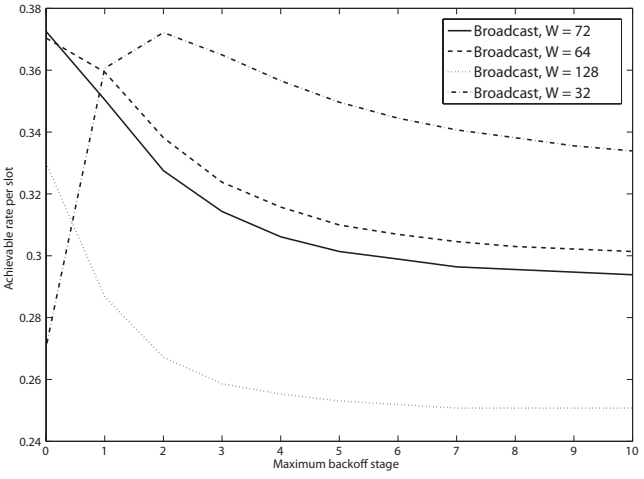

Fig. 4. Broadcast polling efficiency in case of infinite retransmission attempts

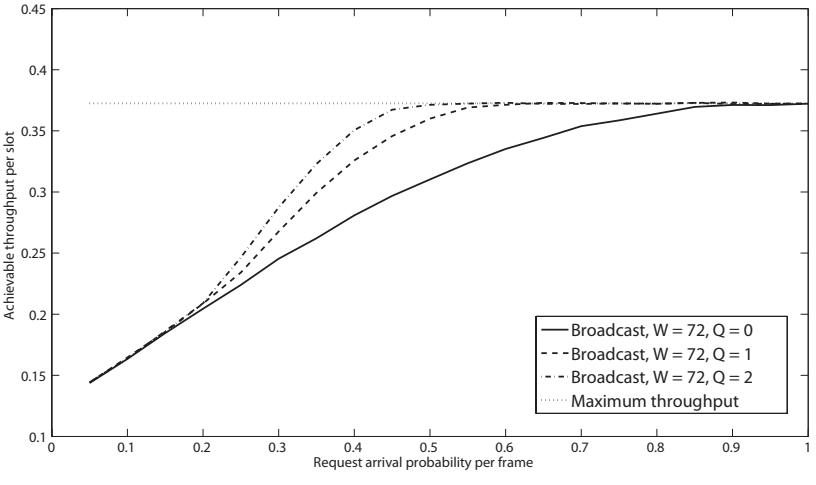

Fig. 5. Broadcast polling lossy throughput performance

5.4 Numerical Results

Here we provide some simulation results that are used to make final conclusions on broadcast and multicast polling efficiency. In Fig. 5 we demonstrate the throughput of the system, where the maximum number of retransmission attempts is set to some natural number, that is, $Q \geq 1$. Therefore, this system represents the intermediate case between those discussed in Section 5.2 and Section 5.3.

We see that for all the values of Q the throughput converges to the value indicated by (18) and (10). However, the convergence is faster for the greater Q value as less requests get discarded. An important conclusion from this is that regardless of the considered system (lossy or lossless) the performance measure of the BEB algorithm is unchangeable, i.e. $T^1_{BEB} = T^{Q+1}_{BEB} \triangleq T_{BEB}$ and $T_{BEB} = R_{BEB}$.

6 Conclusion

In this paper we firstly considered user grouping for the infinite population model and showed that the RMA system capacity remains unchanged. Additionally, we introduced the lossy and lossless system types for which the performance of the BEB algorithm was investigated. Using various analytical techniques, it was demonstrated that the BEB throughput in the lossy system coincides its rate in the lossless one. An important optimization of the BEB parameters was shown and its application in IEEE 802.16 standard was discussed.

The conclusion we make is that multicast polling gain over broadcast polling for the practical scenarios with symmetric grouping, where the groups have equal size and the BEB parameters are the same for each group is minor and decreases as the user population grows. However, to support the QoS requirements, another grouping may be applied, with different BEB parameters and/or unequal group sizes. The contention-based polling performance for these scenarios is subject to a separate investigation, but the demonstrated approaches remain, nevertheless, applicable.

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. 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)
4. Goodman, J., Greenberg, A., Madras, N., March, P.: Stability of binary exponential backoff. Journal of the ACM 35(3), 579–602 (1988)
5. Paterakis, M., Georgiadis, L., Papantoni-Kazakos, P.: On the relation between the finite and the infinite population models for a class of raa's. IEEE Transactions on Communications 35, 1239–1240 (1987)
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. 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)
14. Tsybakov, B., Likhanov, N.: Upper bound on the capacity of a random multiple-access system. Problems of Information Transmission 23(3), 224–236 (1987)
15. Turlikov, A., Vinel, A.: Capacity estimation of centralized reservation-based random multiple-access system. In: Symposium on Problems of Redundancy in Information and Control Systems, vol. 1, pp. 154–160 (2007)
16. Tsybakov, B.: One stochastic process and its application to multiple access in supercritical region. IEEE Transactions on Information Theory 47(4), 1561–1569 (2001)
17. Tsybakov, B., Mikhailov, V.: Random multiple packet access: Part-and-try algorithm. Problems of Information Transmission 16(4), 305–317 (1980)
18. Tsybakov, B., Berkovskii, M.: Multiple access with reservation. Problems of Information Transmission 16(1), 35–54 (1980)
19. Vinel, A., Zhang, Y., Ni, Q., Lyakhov, A.: Efficient request mechanisms usage in ieee 802.16. In: IEEE Global Telecommunications Conference, vol. 1, pp. 1–5 (2006)
20. Abramson, N.: The throughput of packet broadcasting channels. IEEE Transactions on Communications 25(1), 117–128 (1977)
21. Vinel, A., Zhang, Y., Lott, M., Tiurlikov, A.: Performance analysis of the random access in ieee 802.16. In: IEEE 16th International Symposium on Personal, Indoor and Mobile Radio Communications, vol. 3, pp. 1596–1600 (2005)