QoS Criteria in IEEE 802.16 Collision Resolution Protocol

Abdelillah Karouit and Abdelkrim Haqiq
Department of Mathematics and Computer
Hassan 1st University
Faculty of Sciences and Techniques, Settat, Morocco
e-NGN research group, Africa and Middle East
Email: {arkaouit ; ahaqiq}@gmail.com

Luis Orozco Barbosa
Albacete Research Institute of Informatics
University of Castilla La Mancha (UCLM)
02071 Albacete, Spain
Email: luis.orozco@uclm.es

Abstract—In this paper, we study the collision resolution protocol of the IEEE 802.16 standard for metropolitan broadband wireless access systems. The standard specifies the use of a random access scheme based on the slotted binary exponential backoff (BEB) algorithm during the bandwidth request phase. We analyze the distributed choice of the retransmission probabilities used by the BEB protocol. We follow a cooperative team problem approach towards the analysis of the throughput and the delay during the bandwidth reservation process. A Markov chain analysis is used to obtain the optimal retransmission probabilities, the optimal value of the initial backoff window size, \( W_{min} \), the throughput and the expected delays experienced by a subscriber station (SS) for placing its bandwidth requests. The results show that the delay of backlogged messages becomes very large in heavy traffic when our objective is to minimize the expected delay of all transmitted messages. This motivate us to pose a problem of minimizing the expected delay (or maximizing the throughput) subject to constraints on the expected delay of backlogged messages as a multicriterion problem.

Index Terms—IEEE 802.16, Binary exponential backoff, Team problem, Markov chain, QoS Criteria.

I. INTRODUCTION

The IEEE 802.16 WiMAX standard [1] is supposed to play an important role in rapid and ubiquitous adoption of broadband wireless access systems worldwide. In the centralized point-to-multipoint architecture of WiMAX, subscriber stations (SSs) share the uplink to a base station (BS) on demand basis. This means that if a SS requires bandwidth, it informs the BS by means of transferring a request message. The BS scheduler accepts the requests from different SSs and grants them the transmission opportunities in time slots by using some scheduling algorithms, which should take into account the requirement of each SS and the available channel resources. These grants are based on the negotiated quality of service (QoS) agreements. Two main methods to provide transmission opportunities are suggested : centralized polling and contention-based random access. According to the IEEE 802.16, the latter may be implemented in two different ways. The first and most used solution is to give all stations an opportunity to access all available contention slots ; another typical solution is to group some stations together and assign disjoint subsets of all contention slots to each group. Previous results have been obtained to evaluate its performance as well as to propose to vendors some efficient ways to use different available tools and methods. One main line of research is devoted to improving the operation of mesh-topology network based on the IEEE 802.16 standard [4]-[5], where this system is analyzed mainly by means of simulations. An analytical model for CSMA based Point-to-Multipoint (PMP) broadband wireless network was considered in [6]. However, according to the WiMAX standard [1] a mandatory random access method in IEEE 802.16 is based on a truncated BEB without any carrier sense features.

A large number of research papers have focused on the performance evaluation of the various IEEE 802.16 features [7]-[8]-[9]. In particular, the bandwidth requests transmission by a subscriber 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 [10]. 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. The asymptotic behavior of the BEB algorithm has been widely addressed in the literature. In [11] it was shown that the BEB algorithm is unstable in the infinitely-many subscribes stations case. By contrast, [12] shows that the BEB is stable for any finite number of subscribes stations, 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 [13]. 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 [14] and [15].

BEB algorithm within the framework of IEEE 802.16 was first investigated in [16] under a “saturation” condition assumption by using an analytical approach similar to [18]. However, this assumption is too strong since networks do not typically operate in saturation conditions. Most Internet
applications exhibit bursty traffic characteristics. A reasonable model for the fair analysis in more general case of an arbitrary request arrival rate should consider subscribers in tow states (active/inactive). Similar idea was introduced in the analysis of wireless centralized networks, for instance, wireless ATM in [19]. Following this approach, we do not consider the packet transmission phase, but limit our scope to the throughputs and delay analysis of bandwidth request during the reservation process.

In this paper, we undertake a team problem approach. We obtain the optimal retransmission probabilities in the team problem to optimize the two parameters of performances (the throughput and delay of bandwidth request) through an analytical model. The overviews of the IEEE 802.16 protocol and its request mechanisms in Section II. We then briefly describe the BEB algorithm in Section III. In Section IV we formulate and solve the team problem. We then provide some numerical results in Section V. In Section VI we pose and solve the multicriterion problem of minimizing the expected delay subject to constraints on the expected delay of backlogged messages only. Finally we draw our conclusions in Section VII.

II. IEEE 802.16 MAC PROTOCOL

A. FRAME STRUCTURE

Let us consider the network with a PMP architecture, which consists of one BS managing multiple SSs. Transmissions between the BS and SSs are realized in fixed-sized frames by means of time division multiple access (TDMA) / time division duplexing (TDD) mode of operation. The frame structure consists of a downlink sub-frame for transmission from the BS to SSs and an uplink sub-frame for transmissions in the reverse direction. The Tx/Rx transition gap (TTG) and Rx/Tx transition gap (RTG) shall be inserted between the sub-frames to allow terminals to turn around from reception to transmission and vice versa. In the downlink sub-frame the Downlink MAP (DL-MAP) and Uplink MAP (UL-MAP) messages are transmitted by the BS, which comprise the bandwidth allocations for data transmission in both down- link and uplink directions, respectively. Another important management message, which is related to UL-MAP, is called an Uplink Channel Descriptor (UCD). It can be periodically transmitted in the downlink sub-frame. The values of the minimum backoff window, $W_{\text{min}}$, and the maximum backoff window, $W_{\text{max}}$, are defined in this message, which are used by the collision resolution algorithm. The uplink sub-frame contains transmission opportunities scheduled for the purpose of bandwidth requests, in which Bandwidth Request (BW-REQ) messages can be transmitted by the SSs to indicate to the BS their uplink bandwidth allocation requirements. The BS manages the number of transmission opportunities through the UL-MAP message.

B. REQUEST MECHANISM IN IEEE 802.16

Each transmission opportunity may be assigned by the BS either to exactly one subscriber station or to a group of stations. In the first case, a station is provided a so-called unicast opportunity for issuing its BW-REQ. In other words, the BS polls a SS to allow it to transmit the request in a contention-free manner. In the latter case random access algorithm is used by the group of SSs to contend for the common transmission opportunities and resolve possible collisions. As mentioned earlier, the mandatory method of contention resolution, whose inclusion is mandatory by the standard, is based on a truncated binary exponential backoff, with the initial backoff window and the maximum backoff window controlled by the BS. This algorithm is described in detail in the next section. The information whether BW-REQ message is successfully transmitted or distorted (because of collision or noise) is not explicitly transmitted by the BS in the downlink. The standard does not specify how the SSs get to know how outcome of their transmission. It might be based on the correspondence between the amount of the resources assigned to the given SS and the amount of the resources it has asked for in the transmitted BW-REQ message.

III. BINARY EXPONENTIAL BACKOFF

In distributed multiple access, a simple yet effective random backoff algorithm is widely used to avoid collisions. In particular, the truncated binary exponential backoff algorithm adjusts the contention window size dynamically according to the collision intensity. This algorithm steps may be summarized as follows:

Rule 1.1. If a new bandwidth request arrives to a SS in the frame $t - 1$ and this SS 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 to which the SS belongs to. Notice, that in case of broadcast polling the SS 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 SS chooses a number (backoff counter) in the range $\{0, 1, ..., 2^{\text{min}(m,i)W - 1}\}$ 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 SS 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 SS 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, 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 SS determines that the (re)transmission of the last request was successful, it resets the collision counter to \( i = 0 \).

IV. PROBLEM FORMULATION

We use a Markovian model based on [21]-[2]-[8]. We assume that there are a finite number of SSs having a buffer sufficient to store exactly one request. We have set the message size used by all SSs equal to one packet. In terms of bandwidth requesting process, each SS may be either in one of two states: thinking or backlogged. A thinking SS has no message ready for transmission and may generate one during a frame with the probability \( q_a \). 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 a 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 slot the backlogged SS \( i \) attempt to retransmit a BW-Req with probability \( q_r^i \). Similarly to [20], we consider that the probability \( q_r^i \) is given as function of backoff window minimal \( W_{min}, q_r^i = \frac{2}{W_{min} + 1} \).

The transition probabilities of the Markov chain are given by

\[
P(n,n+k)(q_r) = \begin{cases} 
Q_r(1,n)Q_a(k+1,n-1) + 
(1 - Q_r(1,n))Q_a(k,n) & 1 \leq k \leq M - n \\
(1 - Q_r(1,n))Q_a(0,n) + 
Q_r(1,n)Q_a(1,n-1) & k = 0 \\
Q_r(1,n)Q_a(0,n-1) & k = -1 \\
0 & \text{Otherwise}
\end{cases}
\]

A. Expected Delay of Transmitted Messages (E.D.T.M.)

The throughput, defined as the sample average of the number of messages that are successfully transmitted, is given by:

\[
\text{thp}(q_r) = \sum_{n=1}^{M} \pi_n(q_r)[P_{n,n-1}(q_r) + Q_a(1,n-1).Q_r(0,n)] + \pi_0(q_r)Q_a(1,0)
\]

Note: The first term of the equality follows from the fact that if the state at the beginning of the slot is \( n > 0 \) then a departure of a backlogged message occurs during that slot with probability \( P_{n,n-1} \), while the departure of a new arriving message occurs with probability \( Q_a(1,n-1).Q_r(0,n) \). The second term takes into account that if \( n = 0 \) then a new arriving message can be successfully transmitted with probability \( Q_a(1,0) \).

The team problem is therefore given as the solution of the optimization:

\[
\max_{q_r} \text{thp}(q_r) \quad \text{s.c} \quad \begin{cases} 
\pi(q) = \pi(q_r)P(q_r) \\
\pi_n(q_r) \geq 0, n = 0, ..., M \\
\sum_{n=0}^{M} \pi_n(q_r) = 1
\end{cases}
\]

A solution to the team problem can be obtained by computing recursively the steady state probabilities [21], and thus obtain an explicit expression for \( \text{thp}(q_r) \) as function of \( q_r \).

Existence of a solution. The steady state probabilities \( \pi(q_r) \) are continuous over \( 0 < q_r \leq 1 \). Since this is not a close interval, a solution does not exist. However, as we restrict to the closed interval \( q_r \in [\epsilon, 1] \) where \( \epsilon > 0 \), an optimal solution indeed exists. Note also that the limit \( \lim_{q_r \to \epsilon} \pi(q_r) \) exists since \( \pi(q_r) \) is a rational function of \( q_r \) at the neighborhood of zero. Therefore for any \( \delta > 0 \), there exists some \( q_r^* > 0 \) which is \( \delta \)-optimal. \( q_r^* > 0 \) is said to be \( \delta \)-optimal if it satisfies \( \text{thp}(q_r^*) \geq \text{thp}(q_r) - \delta \) for all \( q_r \in [0, 1] \).
In order to calculate the expected delay of the transmitted messages of the team problem, we should first compute the average number of backlogged messages which is given by:

\[ S(q_r) = \sum_{n=0}^{M} \pi_n(q_r)n \]  

(5)

where the throughput of the number of messages that are successfully transmitted is given by 3

The delay of transmitted messages \( D \), defined as the average time slots that a message takes from its source to the receiver can then be obtained by applying Little’s result:

\[ D(q_r) = 1 + \frac{S(q_r)}{thp(q_r)} \]  

(6)

Note that the first term accounts for the first message transmission from the SS.

The team problem E.D.T.M. is therefore given as the solution of the optimization:

\[ \min_{q_r} D(q_r) \quad \text{s.c} \quad \begin{cases} 
\pi(q_r) = \pi(q_r)P(q_r) \\
\pi_n(q_r) \geq 0, n = 0, \ldots, M \\
\sum_{n=0}^{M} \pi_n(q_r) = 1 
\end{cases} \]  

(7)

Existence of a solution The same reasoning as above is valid for any other objective function and in particular herein towards the minimization of the delay of transmission message.

B. Expected Delay of Backlogged Messages (E.D.B.M)

Another relevant quantity in this context is the expected delay of backlogged message \( D^c \) (E.D.B.M) which is defined as the average time, in slots, that a backlogged message takes to go from the SS to the BS. To compute it, we need to calculate the total throughput of backlogged messages which is defined as the sample average of the number of new arriving messages that become backlogged. This is given by

\[ thp^c(q_r) = \sum_{n=0}^{M-1} \sum_{i=1}^{M-n} P_{n,n+i}(q_r)i\pi_n(q_r) \]  

(8)

Applying Little’s result, the expected delay of message that arrive and become backlogged is given by

\[ D^c(q_r) = 1 + \frac{S(q_r)}{thp^c(q_r)} \]  

(9)

The team problem E.D.B.M is therefore given as the solution of the optimization:

\[ \min_{q_r} D^c(q_r) \quad \text{s.c} \quad \begin{cases} 
\pi(q_r) = \pi(q_r)P(q_r) \\
\pi_n(q_r) \geq 0, n = 0, \ldots, M \\
\sum_{n=0}^{M} \pi_n(q_r) = 1 
\end{cases} \]  

(10)

V. NUMERICAL RESULTS

In this section we shall obtain the optimal values of the retransmission probabilities and the optimal size of the backoff window \( W_{\min} \). We further investigate the dependence of the throughput and delay on the arrival probabilities \( q_a \) and on the number of SSs.

A. Expected Delay Of Transmitted Messages

We consider the team problem in which the performances to optimize are the throughput and the expected delay of transmission messages. In Figures 1 and 2 the expected delay of transmitted messages and the optimal retransmission probabilities are plotted versus the arrival probability \( q_a \) for the team problem.
The results show that the optimal retransmission probabilities, $q_r$, should be reduced as the arrival probabilities increase or as the number of SSs is large. In fact, the results show that for arrival probabilities as low as 0.05 and for M=20, the retransmission probability should be reduced drastically.

Figure 3 shows the throughput for $M = 2,4,10,20$ for the team problem, as a function of the arrival probability $q_a$. Under heavy traffic, the throughput decreases drastically as a function of the traffic intensity and much faster in the case of a network consisting of a larger number of SSs. Figure 4 shows the optimal window backoff $W_{min}$ for $M = 2,4,10,20$ and as a function of the arrival probabilities. As expected, when the network is exposed to heavier load, larger number of active SSs or higher arrival rates, the window has to be widen in order to reduce the collision probability.

Fig. 3. Optimal throughput as a function of the arrival probabilities $q_a$ for $M=2, 4, 10, 20$

Fig. 4. The optimal window backoff $W_{min}$ as a function of the arrival probabilities $q_a$ for $M = 2, 4, 10, 20$

Fig. 5. Optimal throughput as a function of the number of SS for $q_a = 0.7, 0.9$

B. Expected Delay of Backlogged Messages

We consider the team problem in which the performance metrics to optimize are the throughput and the expected delay of backlogged messages given by 8 and 10, respectively. We compare the performance for different scenarios: team problem with E.D.T.M and team problem with E.D.B.M. In Figure 7, we plot the throughput using the optimal backoff window size $W_{min}$ for the team problem with E.D.T.M and E.D.B.M versus the arrival probability $q_a$ for $M=10$, respectively. As shown in Figure 7, the throughput of the E.D.B.M. team problem decreases drastically at high arrival probabilities. The backoff window $W_{min}$ remains almost constant as $q_a$ increases, see Figure 4.

Another interesting observation can be made when compar-
VI. QoS CRITERIA

As shown in last section, the delay of backlogged messages become very large in heavy traffic when our objective is to minimize the expected delay of all transmitted messages (or equivalently, when maximizing the throughput). Thus we can distinguish two separate QoS criteria: the total expected delay of transmitted messages as well as the delay of backlogged messages only. We then consider the team problem of minimizing the expected delay of transmitted messages subject to a constraint on the expected delay of backlogged messages. The new team problem that has a new QoS constraint is given by:

$$\min_{q_r} D(q_r) \quad \text{s.t.} \quad D^c(q_r) < d$$

(11)

$d$ is some constant. We shall denote by $d_{max}$ the smallest value of $d$ that insures that there is a solution to 11 for any value of $q_a$. There are indeed values of $d$ for which , for some $q_a$, the problem 11 don’t admit a solution.

Note that due to 6, maximizing the throughput is equivalent to minimizing $S(q_r)$. Hence, we can deduce the fact that the maximizing the throughput is equivalent to minimizing the expected delay. Therefore the retransmission probabilities that solve 11 also solve the problem:

$$\max_{q_r} \text{thp}(q_r) \quad \text{s.t.} \quad D^c(q_r) < d$$

(12)

Figure 8 shows the expected delay of transmitted messages

with and without constraints for $M = 10$. We observe that for all values of $q_a$, the expected delay of transmitted messages without QoS constraint equal to that with QoS constraints, the constraints does not allow the system to decrease the delay at the expense of the backlogged messages, there is no performance degradation in the expected delay of transmitted messages. For $d = 120$, we don’t have any feasible strategy satisfying the QoS for some value of $q_a$, the feasible region.
of strategies is not empty under the condition \( q_a \leq 0.4 \), but for \( d = d_{\text{max}} = 563.5395 \), we have a feasible region for all values of \( q_a \) because \( d_{\text{max}} \) is the maximum expected delay of backlogged messages at optimal retransmission policy for \( q_a \in [0,1] \) which explains the existence at least one feasible strategy for any values of \( q_a \).

VII. CONCLUSION

We have used a game theory approach for setting up the retransmission probabilities and window size used in the collision resolution protocol of the IEEE 802.16 standard. Our objective has been to minimize the expected delay and maximize the throughput during the bandwidth request process. Our results show that as the request arrivals increase, the SSs should reduce the retransmission probability by widening their backoff windows. We have also observed that the delay of the backlogged message rapidly increases under heavy traffic load conditions, and the new constraint does not allow the system to decrease the delay at the expense of the backlogged messages, there is no performance degradation in the expected delay of transmitted messages.

REFERENCES