Overall Delay Analysis of IEEE 802.16 Network

Sergey Andreev
State University of Aerospace
Instrumentation (SUAI), Russia
serge.andreev@gmail.com

Zsolt Saffer
Budapest University of Technology
and Economics (BUTE), Hungary
safferzs@hit.bme.hu

Alexey Anisimov
H&NM Motorola
Software Organization, Russia
alexey.anisimov@motorola.com

***Abstract*—In this paper we conduct a delay analysis of IEEE 802.16 wireless metropolitan area network. In particular, we address the overall message delay, which consists of the reservation and scheduling components. Unicast polling is used for bandwidth reservation and round-robin scheduling is applied at the base station. A discrete-time analytical model is developed with general independent and identically distributed arrivals during a time slot. The model enables asymmetric traffic flows and different message sizes at the subscriber stations. The exact mean overall delay is obtained for the nrtPS service flow in the scenario when the base station splits the subscriber stations into individually polled groups. The analytical model is verified by means of simulation.**

Keywords: IEEE 802.16, WMAN, performance evaluation, bandwidth reservation, polling, queuing model, overall delay.

I. Introduction

IEEE 802.16 standards family 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. The most recent IEEE 802.16e [1] standard is likely to emerge as an outstanding cost-competitive technology mainly due to its long range and sophisticated Quality-of-Service (QoS) support.

Despite the number of enhancements to the first version of IEEE 802.16 standard, a lot of critical issues are left out of its scope. Among them are exact scheduling algorithms at both base station and subscriber station and methods to ensure QoS requirements of the end users. Therefore, a lot of research papers address these problems. In [2], [3] and [4] various frameworks are built and analyzed to guarantee a specified level of QoS. Efficient bandwidth requesting mechanisms are also discussed in the literature. For instance, in [5], [6] and [7] existing polling schemes are studied as well as prominent modifications of them are considered.

There are also more general works on various analytical approaches to analyze multiple access systems, like the fundamental papers [8] and [9]. But they do not provide any practical application of the considered models. On the other hand, some authors consider various techniques to address the practical performance measures of IEEE 802.16 system. For example, in [10] and [4] the overall system delay is estimated and verified. However these methods are complicated and result only in a rough approximation of the overall delay. In this paper, which is a continuation of our previous work [11], we establish an exact analytical model for the IEEE 802.16 overall delay. Our analytical approach applies a discrete-time GI/D/1 queue with vacations as a special case of the cyclic service system [9] to describe the IEEE 802.16 performance.

The rest of the paper is structured as follows. Section II gives a brief overview of IEEE 802.16 MAC. In Section III we provide the description of the system model and notations. We conduct the mean delay analysis in Section IV. The symmetric load situation as an important special case is treated in Section V. The verification of the analytical results follows in Section VI. Finally, we give a summary in Section VII.

II. Overview of IEEE 802.16

A. MAC layer

The basic point-to-multipoint (PMP) IEEE 802.16 architecture (see Figure 1) assumes that there are one Base Station (BS) and one or more Subscriber Stations (SSs). The packets are exchanged between BS and SSs via separate channels. A DownLink (DL) channel is used for the traffic from the BS to the SSs and the UpLink (UL) channel is used in the reverse direction. There is no need for multiple access on the DL channel, while the UL channel is shared among the SSs. The access procedure to the UL channel is the subject to one of the specified multiple access protocols.

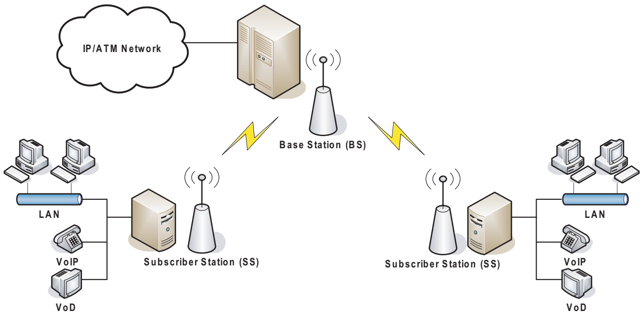

Fig. 1. An Example for IEEE 802.16 PMP Architecture.

The standard defines two mechanisms of multiplexing DL and UL channels: Time Division Duplex (TDD) and Frequency Division Duplex (FDD). In TDD mode the frame is divided between the DL part and the UL part. The border between these parts may change dynamically depending on the SSs bandwidth requirements. The SSs access the UL channel by means of Time-Division Multiple Access (TDMA). The structure of the MAC frame in TDD/TDMA mode is shown in Figure 2. In FDD mode the SSs transmit in different subbands.

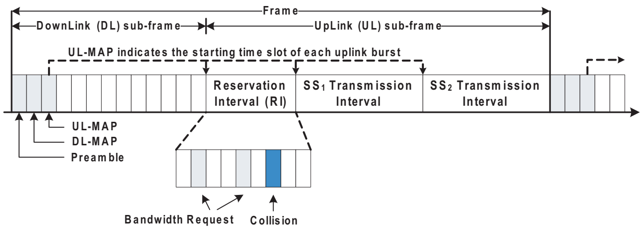
Frame
DownLink (DL) sub-frame
UpLink (UL) sub-frame
UL-MAP indicates the starting time slot of each uplink burst
Reservation Interval (RI)
SS$_1$ Transmission Interval
SS$_2$ Transmission Interval
UL-MAP
DL-MAP
Preamble
Bandwidth Request
Collision

Fig. 2. IEEE 802.16 MAC frame structure in TDD/TDMA mode.

In the DL channel the BS - as the only sending station - broadcasts the packets to all the SSs. Together with the data packets, the BS also transmits service information about the slots which are allocated for each of the SSs in the UL channel. This information is incorporated in the UL-MAP message and is used by the SSs for scheduling their data packets in the UL channel. To allow feedback from the SSs the BS also specifies a portion of channel resources as the Reservation Interval (RI). During the RI the SSs transmit their bandwidth requests (BW-Req), which are then processed by the BS.

The access procedure of the SSs to the RI could be either contention-based or contention-free. The latter is referred to as unicast polling and corresponds to the case when BS assigns to each of the SSs a transmission opportunity for its bandwidth request. The former comprises two mechanisms, namely, multicast and broadcast polling. When broadcast polling is enabled all the SSs are expected to send their bandwidth requests by choosing one of all the transmission opportunities uniformly. During the access to the RI request collisions may occur, which may be subject to a subsequent resolution. The specified collision resolution algorithm is the truncated binary exponential backoff. In case of multicast polling the SSs are polled in groups and within a group the rules of broadcast polling are applied. Additionally, IEEE 802.16 enables piggybacking for sending BW-Reqs attached to data packets.

B. *Service flow types*

IEEE 802.16 was designed to support a variety of traffic types. It should be efficient for high data rate applications (video streaming) as well as for low data rate applications (web surfing). IEEE 802.16 effectiveness should not degrade in case of bursty traffic and delay-critical applications (voice over IP (VoIP), audio). The main challenge in ensuring QoS requirements in IEEE 802.16 is that all the traffic types with respective characteristics should be serviced at the same time. For this purpose the standard defines five QoS classes, which are described below.

1) Unsolicited Grant Service (UGS) is oriented at the real-time traffic where fixed-size data packets are generated periodically (CBR input source).
2) Real-Time Polling Service (rtPS) is oriented at the real-time traffic where variable-size data packets are generated periodically (VBR input source).
3) Non Real-Time Polling Service (nrtPS) is like rtPS but data packet generation is not necessarily periodic.
4) Best Effort (BE) is suitable for applications where no throughput or delay guarantee is provided.
5) Extended Real-Time Variable Rate (ERT-VR) is like rtPS but with more strict delay requirement (guaranteed jitter) to support real-time applications like VoIP with silence suppression. This class is defined only in the recent IEEE 802.16e [1] standard and is often referred to as Extended Real-Time Polling Service (ertPS).

III. Model and notations

It follows from [1] that unicast polling of SSs is the most common bandwidth reservation mechanism in the UL channel. Hence in this paper we concentrate on this mechanism only. We assume that there are only two flows of different types at one time instant in the system. We consider one prioritized service flow in the system (nrtPS), while the remaining bandwidth is utilized by the non-prioritized flow (BE). The prioritized service flow can be extended by the rtPS and ertPS QoS classes, in which case the overall delay analysis could help to design the enabled traffic for these classes.

A. *Restrictions of the model*

Our overall delay model considers 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 flow types are allowed, however we address the delay of only the prioritized flow type.

R.4 The used bandwidth reservation mechanism is unicast polling.

R.4 The uplink scheduler applies a round-robin (RR) scheduling.

R.6 One connection per SS is allowed.

R.7 Piggybacking is not used.

B. *General scenario*

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. We apply a discrete-time model, in which the time is slotted. Each slot is equal to the transmission time of a data packet. However, all the time durations are measured between the relevant events in real-value. This slight modification of the discrete-time model ensures that the analytical results fit those of the corresponding continuous-time model. Messages transmitted by the SSs consist of a number of data packets.

C. *Grouping of SSs*

We limit the split-up of the maximum available duration of the uplink data transmission. This can be used to control the bandwidth usage among the prioritized and the non-prioritized service flows. In one frame only $P \leq N$ SSs are allowed to

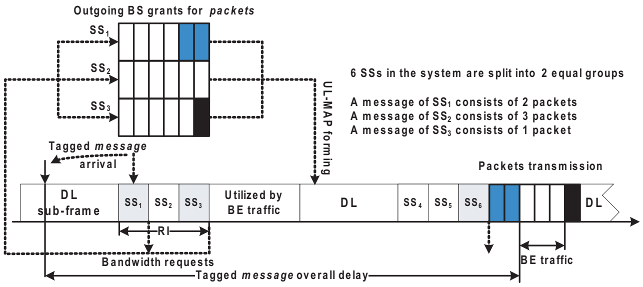
Outgoing BS grants for *packets*
SS_1
SS_2
SS_3
6 SSs in the system are split into 2 equal groups
A message of SS_1 consists of 2 packets
A message of SS_2 consists of 3 packets
A message of SS_3 consists of 1 packet
UL-MAP forming
Tagged *message* arrival
Packets transmission
DL sub-frame
SS_1
SS_2
SS_3
Utilized by BE traffic
DL
SS_4
SS_5
SS_6
DL
RI
Bandwidth requests
BE traffic
Tagged *message* overall delay

Fig. 3. Grouping and scheduling model example.

transmit on the uplink, each of them only one message. In the next frame the next *group* of P SSs is allowed to transmit, and so on. Thus the BS divides all the SSs into groups of P SSs. The unicast polling is performed also on periodic basis. In one frame only SSs belonging to one group are polled and are allowed to send a BW-Req. Then the individual groups are polled in consecutive frames. Clearly, the number of groups is:

$$L = N/P. \tag{1}$$

We call the i-th SS of the j-th group as SS $i(j)$ and the messages arriving to it as $i(j)$-messages ($i = 1, \ldots, P$, and $j = 1, \ldots, L$).

We denote the duration of each frame by T_f. Clearly, L consecutive frames constitute a cycle, which lasts for:

$$C = LT_f. \tag{2}$$

D. RR Scheduling

A BW-Req sent by a SS $i(j)$ represents the request for all $i(j)$-messages, which are accumulated in its outgoing buffer in the last cycle, i.e., since its last BW-Req sending.

For each SS the BS maintains an individual buffer with infinite capacity. Let $i(j)$-*polling slot* stands for the i-th polling slot in the RI of the frame, in which the j-th group is polled. For each group j the BS performs an immediate processing at the end of each $i(j)$-polling slot in the order of polling of the SSs of group j. If BW-Req is received from SS $i(j)$, then first a BS grant is assigned to each data message represented by the request and these BS grants are put into the corresponding individual BS grant buffer of SS $i(j)$ (according to their order in the request). At the end of the $i(j)$-polling slot - after a potential BW-Req handling from SS $i(j)$ - the BS takes the first BS grant (if there is any) from the individual grant buffer of SS $i(j)$ and schedules it for sending in the UL-MAP of the next frame. Hence in case of finding the individual BS grant buffer of SS $i(j)$ empty upon receiving a BW-Req of an $i(j)$-message its data is transmitted uplink in the frame next to the one, in which the BW-Req is received. The BS processing is illustrated in Figure 3.

This way the BS realizes RR scheduling via the intra group BS processing at the end of the polling slots of the same frame and via the periodic BS processing of groups in consecutive frames.

In the case when one or more SSs have no message of the prioritized service flow 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 the scope of this paper.

E. Analytical approach

The numbers of arriving $i(j)$-messages during each slot are assumed to be independent, identically distributed random variables, and thus the numbers of arrivals in different slots are also independent of each other. The duration of a transmission slot is τ. We express the numbers of arrivals in messages per time unit. The first and second moments of the number of arriving $i(j)$-messages per time unit are denoted by $\lambda_{i,j}$ and $\lambda_{i,j}^{(2)}$, respectively. Hence the overall arrival rate is $\lambda = \sum_{j=1}^{L} \sum_{i=1}^{P} \lambda_{i,j}$. The messages are assumed to be of fixed length. $b_{i,j}$ denotes the size of an $i(j)$-message, i.e., the number of packets (slots) in a message arriving to station $i(j)$. The arrival processes and the message sizes (in slots) at the different stations are assumed to be mutually independent.

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

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

The transmission time of a BW-Req is α. Hence, $T_{pi} = P\alpha$ and we get:

$$T_{ud} = T_u - P\alpha. \tag{3}$$

For each group the available duration of the uplink data transmission in a frame is T_{ud}. Then for the capacity allocation of the SSs for each group we obtain:

$$T_{ud} = \sum_{i=1}^{P} b_{i,j}\tau, \quad j = 1, \ldots, L. \tag{4}$$

F. Model assumptions

We denote the *utilization* of SS $i(j)$ by $\rho_{i,j}$. Since each SS gets a chance to transmit on UL at most one message in each L consecutive frames, we get for the utilization of of SS $i(j)$:

$$\rho_{i,j} = \lambda_{i,j} LT_f. \tag{5}$$

Additionally, we formulate the following assumptions of our model:

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

$$\rho_{i,j} = \lambda_{i,j} LT_f < 1, \;\; i = 1, \ldots, P, \;\; j = 1, \ldots, L. \tag{6}$$

This relation ensures the stability of the model.

A.2 The time of BS processing and scheduling are negligible.

A.3 The channel propagation time is negligible.

A.4 The transmission channels are error-free.

IV. Overall delay analysis

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

A. Overall delay definition

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

$$W_{i,j} = W_{i,j}^r + \alpha + W_{i,j}^s + W_{i,j}^t + b_{i,j}\tau, \qquad (7)$$

where $W_{i,j}^r$ is the reservation delay, which is defined as the time interval from the $i(j)$-message arrival to SS $i(j)$ until the start of sending a corresponding BW-Req to the BS.

α is the transmission time of a BW-Req.

We define the *grant time of the tagged $i(j)$-message* as the end of the $i(j)$-polling slot in the frame preceding the one, in which the tagged $i(j)$-message is transmitted.

$W_{i,j}^s$ is the scheduling delay, which is defined as the time interval from the end of sending a BW-Req of the tagged $i(j)$-message to its grant time.

$W_{i,j}^t$ is the transmission delay, which is defined as the time interval from the grant time of the tagged $i(j)$-message to the start of its successful transmission in the UL sub-frame.

$b_{i,j}\tau$ is the transmission time of an $i(j)$-message.

B. 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. As the frame length, and hence the cycle length is fixed, 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.

An $i(j)$-message arriving during an $i(j)$-polling slot must wait with BW-Req sending until the next $i(j)$-polling slot (cycle time C). Therefore the start epochs of the $i(j)$-polling slots divide the time axis into $C = LT_f$ fixed-length cycles. This epoch triggers α time later either a BS grant scheduling for this SS (which results in an uplink $i(j)$-message transmission in the next frame), or not, in which case no BW-Req received from this SS during the actual $i(j)$-polling slot and there is no waiting BS grant in the individual BS buffer of this SS. These events qualify the cycles as "service" or "vacation". Note that the event of a BS grant scheduling is exactly a grant time of an $i(j)$-message. Therefore the waiting time of this queueing model consists of the reservation and scheduling delays at SS $i(j)$ together (without the transmission time of a BW-Req).

Hence the appropriate queuing system to describe the behavior of SS $i(j)$ is an GI/D/1 queue with vacation, in which both the service time and the vacation time are deterministic and equal to $C = LT_f$. Therefore, we apply the discrete-time mean waiting time formula of a cyclic service system [9] to our vacation model with the corresponding parameters. Due to the slight modification in our discrete-time model the waiting time of an $i(j)$-message starts at its arival epoch instead of the next slot boundary. Taking it into account and using also (1) and (2), the mean of the sum of the reservation and scheduling delays of the tagged $i(j)$-message can be expressed as:

$$\begin{aligned} & E\left[W_{i,j}^r + W_{i,j}^s\right] = \qquad (8) \\ & \frac{C}{2(1-\lambda_{i,j}C)} + \frac{\left(\lambda_{i,j}^{(2)} - \lambda_{i,j}^2 - \lambda_{i,j}\right)C}{2\lambda_{i,j}(1-\lambda_{i,j}C)} \\ & = \frac{NT_f}{2(P-\lambda_{i,j}NT_f)} + \frac{\left(\lambda_{i,j}^{(2)} - \lambda_{i,j}^2 - \lambda_{i,j}\right)NT_f}{2\lambda_{i,j}(P-\lambda_{i,j}NT_f)}. \end{aligned}$$

C. Transmission delay

The transmission delay is a fixed time from the end of the $i(j)$-th polling slot to the start of transmission of the tagged $i(j)$-message in the UL sub-frame of the next frame:

$$\begin{aligned} W_{i,j}^t &= T_f - \alpha i - T_d + T_d + P\alpha + \sum_{k=1}^{i-1} b_{k,j}\tau \\ &= T_f + \alpha(P-i) + \sum_{k=1}^{i-1} b_{k,j}\tau. \qquad (9) \end{aligned}$$

D. Mean overall delay

Accounting for (7), the mean overall delay is given by:

$$E\left[W_{i,j}\right] = E\left[W_{i,j}^r + W_{i,j}^s\right] + \alpha + E\left[W_{i,j}^t\right] + b_{i,j}\tau.$$

Substituting the expressions (8) and (9) we obtain the expression of the mean overall delay:

$$\begin{aligned} E\left[W_{i,j}\right] = & \frac{NT_f}{2(P-\lambda_{i,j}NT_f)} + \frac{\left(\lambda_{i,j}^{(2)} - \lambda_{i,j}^2 - \lambda_{i,j}\right)NT_f}{2\lambda_{i,j}(P-\lambda_{i,j}NT_f)} \\ & + T_f + \alpha(P-i+1) + \sum_{k=1}^{i-1} b_{k,j}\tau + b_{i,j}\tau. \qquad (10) \end{aligned}$$

V. Symmetric load situation

Here we consider an important special case of the above grouping system, i.e. when SSs have symmetric loads and values $b_{i,j}$ are equal. For this special case we set up a second analytical model to cross-check our general model presented in Section IV. We assume for simplicity that $b_{i,j} = 1$, that is the number of packets per message is equal to one transmission slot. This makes (4) to take the form of $T_{ud} = P\tau$.

We are interested in the analysis of the message overall delay W in the described system. From a viewpoint of a particular SS the observed grouping system could be described by a GI/D/1 queueing model, since service time in the BS queue is deterministic and equal to $C = LT_f$. This is explained by the fact that each SS has a chance to transmit its message (if any) once in a polling cycle. We assume Poisson input flow of messages into the system. This makes the M/D/1 queueing model to be the analytical approach to the considered system model.

Following the approach of [12] it is very convenient to use the notion of the residual service time to prove the following simple formula:

$$E[W] = E[W_0] + \frac{\overline{X}\rho_i}{2(1-\rho_i)}, \tag{11}$$

where W_0 is message overall delay conditioning on the fact that the arriving message sees the system empty, $\overline{X}$ is the service time and ρ_i is the corresponding utilization factor for the SS number i. Notice, that the second index j may be omitted in this section, as all the SSs are identical.

The service time for the particular user was explained above to be $\overline{X} = C = LT_f$. The utilization factor is defined to be $\lambda_i\overline{X}$, where λ_i is the message arrival rate for one SS (here we also omit index j). Since the considered system is symmetric, we notice that $\lambda_i = \frac{\lambda}{N}$, where λ is the overall arrival rate and, consequently:

$$\rho_i = \frac{LT_f\lambda}{N}. \tag{12}$$

Consider now W_0, that is, the message overall delay in the empty system. W_0 is defined as the sum of the following components:

$$W_0 = W^r + W^s + W^t, \tag{13}$$

where W^r is the reservation delay, W^s is the scheduling delay and W^t is the transmission delay.

W^r is defined as the time interval from the moment of the message arrival at the SS to the end of the RI of a frame, in which this SS was polled. Thus W^r includes also the transmission time of a BW-Req. The scheduling delay is defined as the time interval from the end of the reservation delay to the end of the RI of the frame, in which this SS transmits. The transmission delay is defined as the time interval from the end of scheduling delay to the end of the actual transmission. Thus the transmission delay includes also the transmission time of the message.

Consider now the components of W_0 individually. Following the approach of [11] it can be shown that

$$E[W^r] = \frac{LT_f + \alpha(P+1)}{2}. \tag{14}$$

Once a message arrives at the empty BS queue, it is transmitted in the next frame according to the scheduling rules above. Hence the scheduling delay is exactly $W^s = T_f$.

To determine W^t assume in the grouping of SSs that each SS have a fixed position in the uplink subframe, that is, if an SS with the lower number has no messages to transmit, there is a gap in transmission. This gap is utilized by the BE traffic (see discussion above). Therefore, $W^t(i) = i\tau$. By averaging over all SS positions in a group, we get $E[W^t] = \frac{P+1}{2}\tau$. Combining the above, we establish the following form of the mean initial delay $E[W_0]$:

$$E[W_0] = \frac{LT_f + \alpha(P+1)}{2} + T_f + \frac{P+1}{2}\tau. \tag{15}$$

Substituting (15) into (11) and after some simplification, we obtain the following closed-form expression for the mean message overall delay $E[W]$:

$$E[W] = T_f + \frac{(P+1)(\alpha+\tau)}{2} + \frac{NT_f}{2(P-\lambda T_f)}. \tag{16}$$

As expected (16) is the special case of (10) since for Poisson arrivals it is known that $\lambda_i^{(2)} = \lambda_i^2 + \lambda_i$.

VI. 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 (see Section III.A). Following [13] we set the simulation parameters of IEEE 802.16 MAC and PHY as follows:

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

TABLE I
Basic IEEE 802.16 simulation parameters

For the purposes of simplicity we restrict our explorations to the symmetric load case. In other words, each message comprises exactly one data packet, that is, all $b_{i,j} = 1$. This results in a better visibility of the below comparison results.

A. Poisson arrivals

First we validate the analytical model developed in Section V for the special case of symmetric load. We compare the behavior of our basic grouping system with a system applying an optimal scheduler. This scheduler leaves no gaps during its operation. For instance, it could serve the arriving messages on a 'first-come-first-serve' basis. For this discipline the shared message buffer at the BS suffices.

We remark that the analytical description of the optimal scheduling system is intractable. Therefore, in Figure 4 we compare the analytical results for the basic grouping system with the simulation results for the optimal scheduling one. For the considered scenario with $N = 20$ and $P = 5$ we observe a fairly good upper bound on optimal overall delay given by our basic grouping system. Clearly, this bound becomes less precise as the arrival rate of messages into the system increases. However, the critical arrival rate is the same for both grouping systems and is equal to $\frac{N}{LT_f}$ for the scenario considered in Figure 4.

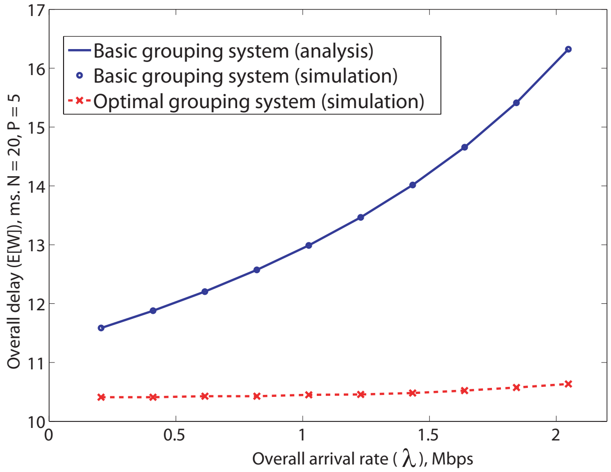

Fig. 4. Grouping systems comparison.

B. Bernoulli batch arrivals

To verify the accuracy of our general analytical approach (Section IV) we show an illustrative example with Bernoulli batch arrival process. We set $\tau = 1$. It follows that the number of arrivals during a slot equals the number of arrivals per time unit. We assume that each SS transmits an aggregated message flow from n end users. Each user generates new message per discrete time slot with some probability p. Therefore, the arrival rate of new messages into the system from one SS is $\lambda_i = np$ and overall arrival rate is $\lambda = N\lambda_i = Nnp$. It is easy to show that the second moment of the number of arriving messages per time unit is equal to $\lambda_i^{(2)} = np(1-p^2) + n^2p^2$. Substituting the corresponding parameters into (10) we obtain the mean overall delay and verify it in Figure 5.

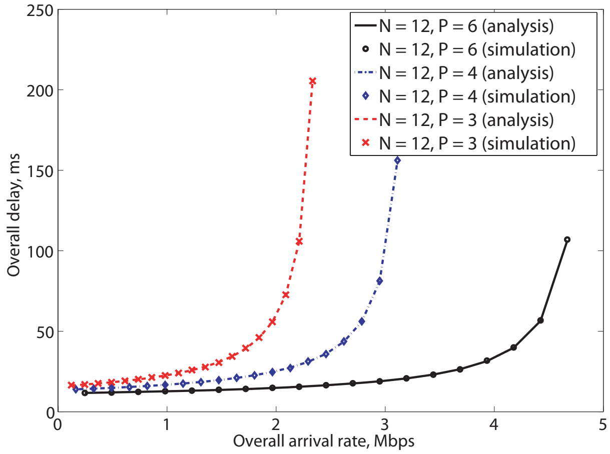

Fig. 5. Different groupings comparison.

We firstly observe the excellent accordance of analytical and simulation results even under near-critical arrival rates. The below curve (corresponding to $N = 12$, $P = 6$) demonstrates the best sustained arrival rate (throughput) and mean delay performance as the cycle length is the shortest of the three groupings considered. However, this results in the full utilization of UL sub-frame and, consequently, oppression of the non-prioritized service flow. The BS may choose to decrease the available throughput by lengthening the cycle ($N = 12$, $P = 4$ and $N = 12$, $P = 3$, respectively). In this case the mean delay of the considered service flow increases, but the available throughput for the transmission of the non-prioritized service flow also grows.

VII. Summary

In this paper we developed an analytical discrete-time model to evaluate the mean overall delay of IEEE 802.16 wireless network. This model enables asymmetric traffic flows and general discrete probability distribution of number of message arrivals per time slot. The model applies unicast polling and round-robin scheduling. We considered the prioritized (nrtPS) and non-prioritized (BE) service flows and established a closed-form expression for the mean overall message delay for the prioritized flow.

We assumed that the BS splits SSs into groups to poll them individually. Simulation results are presented that show the accuracy of our model. On one hand the grouping of SSs results in a longer polling cycle and, consequently, higher delay. On the other hand, longer cycle with constant frame duration leaves more throughput available to the non-prioritized flow in the system. Therefore our approach gives BS a mechanism to trade between the throughput of the non-prioritized service flow and the delay of the prioritized service flow.

References

[1] *IEEE Std 802.16e-2005, Piscataway, NJ, USA, December 2005.*

[2] G.S. Paschos, I. Papapanagiotou, C.G. Argyropoulos, and S.A. Kotsopoulos. A heuristic strategy for ieee 802.16 wimax scheduler for quality of service. In *45th Congress FITCE*, 2006.

[3] L.F.M. de Moraes and P.D. Maciel. A variable priorities mac protocol for broadband wireless access with improved channel utilization among stations. *Int. Telecomm. Symp.*, 1:398–403, 2006.

[4] Y.-J. Chang, F.-T. Chien, and C.-C. J. Kuo. Delay analysis and comparison of ofdm-tdma and ofdma under ieee 802.16 qos framework. *IEEE Global Telecomm. Conf. (GLOBECOM)*, 1:1–6, 2006.

[5] X. Wang, Y. Yu, and G.B. Giannakis. Combining random backoff with a cross-layer tree algorithm for random access in ieee 802.16. *IEEE Wireless Comm. and Networking Conf. (WCNC)*, 2:972–977, 2006.

[6] A. Vinel, Y. Zhang, Q. Ni, and A. Lyakhov. Efficient request mechanism usage in ieee 802.16. *IEEE Global Telecomm. Conf. (GLOBECOM)*, 1:1–5, 2006.

[7] O. Alanen. Multicast polling and efficient voip connections in ieee 802.16 networks. *10th ACM Symposium on Modeling, analysis, and simulation of wireless and mobile systems*, 1:289–295, 2007.

[8] I. Rubin. Access-control disciplines for multi-access communication channels: Reservation and tdma schemes. *IEEE Trans. Inf. Theory*, 25(5):516–536, 1979.

[9] O.J. Boxma and W.M. Groenendijk. Waiting times in discrete-time cyclic-service systems. *IEEE Trans. on Comm.*, 36(2):164–170, 1988.

[10] R. Iyengar, P. Iyer, and B. Sikdar. Delay analysis of 802.16 based last mile wireless networks. *IEEE Global Telecomm. Conf. (GLOBECOM)*, 5:3123–3127, 2005.

[11] Z. Saffer and S. Andreev. Delay analysis of ieee 802.16 wireless metropolitan area network. In *Int. Workshop on Multiple Access Communications (MACOM)*, 2008.

[12] D. Bertsekas and R.Gallager. *Data Networks*. Prentice-Hall, 1st ed., 1987; 2nd ed., 1992.

[13] D. Sivchenko, N. Bayer, B. Xu, V. Rakocevic, and J. Habermann. Internet traffic performance in ieee 802.16 networks. In *European Wireless*, 2006.