The Synergy of Broadcast Polling and Piggybacking in WiMAX Networks

Maria Iloridou, Evangelos Papapetrou, *Member, IEEE*, and Fotini-Niovi Pavlidou *Senior Member, IEEE*

Technical Report TR-2016-2

December 9, 2016

***Abstract*—In this work we present a model for analyzing the combined use of broadcast polling and piggybacking in Worldwide Interoperability for Microwave Access (WiMAX) networks. For an accurate analysis of piggybacking, the model focuses on the realistic case of limited up-link bandwidth and non-trivial queueing capability at the subscriber stations. We first model the activity of a subscriber station using a Markov chain and its queue as an M/G/1 system with vacations in order to facilitate the analysis of the piggyback mechanism. We then derive a set of fixed point equations that describe not only the contention process at the network level but also bandwidth allocation to contending and piggybacked requests. Our model uses a minimal set of assumptions and is generic in the sense that it is customizable through a set of parameters. It can also reproduce the system performance in both saturated and non-saturated conditions. After validating our analysis through extensive simulations, we shed light on the aspects of the synergy between broadcast polling and piggybacking and unveil the pros and cons of using the latter.**

***Index Terms*—piggyback, broadcast polling, IEEE 802.16**

I. Introduction

WORLDWIDE Interoperability for Microwave Access (WiMAX) is probably the most promising broadband wireless technology for the next decades. WiMAX has been standardized by the IEEE 802.16 Working Group [1]. To meet the service requirements of various applications the *medium access control (MAC)* sublayer of the IEEE 802.16 standard supports five scheduling services in the popular *Point-to-Multipoint (PMP)* mode of operation: the unsolicited grant service (UGS), the real-time Polling Service (rtPS), the extended real-time Polling Service (ertPS), the non-real-time Polling Service (nrtPS) and the Best Effort (BE) one. Among these services BE holds a central role not only due to its inherent compatibility with the Internet architecture but also because it is the choice for a variety of WiMAX application classes such as media content downloading, web browsing, instant messaging, etc [2]. According to the standard, the mandatory medium access method for BE follows a contention-based approach that features a bandwidth request (BR)-grant scheme. In other words, subscriber stations (SSs) with BE data to transmit should contend for receiving a bandwidth (BW) grant from the base station (BS).

In this work we focus on the realization of the contention-based access scheme for BE in the widespread scenario of fixed access networks, also known as *broadcast polling* [1]. Analyzing this mechanism is of paramount importance not only due to the key role of BE but also because it is used by other scheduling services such as nrtPS and ertPS [1], [2]. Reasonably, several researchers looked into analyzing broadcast polling [3]–[12] (as well as its variant, group/multicast polling [13], [14]) and provided valuable insight into its characteristics. Yet, SSs implementing broadcast polling may also use on top of it the optional *piggyback* mechanism. Albeit the latter is a cost-effective alternative that can provide significant performance improvements [7], [15], so far its analysis has received little attention [7], [8], [16], [17]. The proposed models are a first step towards evaluating piggybacking but unfortunately cannot accurately capture all of its performance features. There are two essential reasons for this, besides the ones relating to the modeling of the underlying contention mechanism (we discuss this latter issue in Section II-B). First, the proposed models assume either no queueing capability [7] or a trivial one (just data from a single arrival) [8]. Consequently, it is not possible to model piggybacking when it is mostly needed, i.e. when an SS's queue builds up due to data arrival bursts. The second major drawback is that unlimited up-link (UL) BW is assumed [8], [16], [17]. This makes it impossible to model the coupling between piggybacking and the contention mechanism that results from sharing UL BW.

Motivated by these observations we focus on analyzing broadcast polling with piggybacking in a setting with limited UL BW, i.e. we wish to model not only the transmission of BRs but also the BW allocation phase. To this end, we follow a typical approach used in the analysis of broadcast polling [6], [12], i.e to model an SS's activity with a Markov chain. Then, we capitalize on concepts from Bianchi's seminal work [18] on IEEE 802.11 networks. Nonetheless, we generalize the Markov chain so as to include a branch that models the piggyback process. More importantly, to facilitate an accurate analysis of piggybacking we consider SSs with non-trivial

M. Iloridou and F.-N.Pavlidou are with the Department of Electrical & Computer Engineering, Aristotle University of Thessaloniki, Thessaloniki, 54636, Greece (e-mail:iloridou@auth.gr,niovi@auth.gr)

E.Papapetrou is with the Department of Computer Science & Engineering, University of Ioannina, Ioannina, 45110, Greece (e-mail:epap@cse.uoi.gr)

queueing capability which we model using an M/G/1 system with vacations. In summary our contributions are:

- We propose a customizable analytical model for broadcast polling with piggybacking (Sections IV and V). The analysis covers both BR transmission and BW allocation phases. It also features a set of realistic characteristics that are critical for modeling piggybacking such as limited UL BW and SSs with non-trivial queueing capability.
- Our analysis is also suitable for investigating plain broadcast polling, i.e. without piggybacking. Compared to other approaches in the literature, our model provides a more detailed modeling of the BW management mechanism (e.g. time window for serving a BR, number of SSs served per frame, etc).
- We provide a detailed performance evaluation of piggybacking using both simulation and analytical results (Section VI).
- We confirm that piggybacking can potentially bring performance improvements but we find that this depends on the ratio of BW allocated for contention to that allocated for data transmission. We discuss the optimal strategy and at the same time we shed light on the trade-offs involved in using piggybacking (Section VI).

As a final note, we believe that our analysis could be a valuable tool when examining piggybacking over the contention-based access scheme for mobile access WiMAX networks since the latter shares striking similarities with broadcast polling.

In the rest of the paper, we provide a brief overview of the contention-based access mechanism for fixed WiMAX networks (Section II-A) and review the related literature (Section II-B). In Section III, we present the system model and the assumptions considered in our analysis. We conclude this work in Section VII.

II. Related Work

A. Overview of Contention-based Access with Piggyback

As mentioned previously, in PMP WiMAX networks, the mandatory access scheme for SSs with BE connections uses a BR-grant schema in a contention-based mode. In other words, an SS, wishing to receive a BW grant for sending data, should first contend for sending to the BS a BR with its BW needs. Regarding the contention process, there are two realizations; one proposed for SC and OFDM physical layer (PHY) specifications, i.e. for fixed access networks, and another one for use with OFDMA PHY utilized for mobile access. The two implementations share many similarities. In the OFDMA case, SSs contend through the transmission of a code in the ranging region of a frame (*Contention-based CDMA BRs*). On the other hand, in fixed access networks all SSs (*broadcast polling*) or a group of them (*group polling*) contend by transmitting BRs in a period of time allocated by the BS in each UL subframe.

In the case of fixed access networks, the contention period is organized in transmission opportunities (TOs), the realization of which depends on the PHY specification. An SS with data to send should wait for a random number of TOs before sending the BR to the next one. This number is uniformly selected from the interval [0,W_0-1], where W_0 is known as the initial contention window size. Note that the waiting period may span multiple frames depending on W_0. If two or more SSs choose the same TO to transmit a BR then a collision occurs. Upon correctly receiving a BR, i.e. no collision occurs, the BS should inform the SS about a BW grant through the UL-MAP in the DL subframe. Since BW is not always available, the BS may provide a grant not necessarily in the next frame but within a number of frames following the one that the successful BR transmission took place. This number is known as the Contention-based Reservation Timeout or the T16 period. The standard does not specify neither carrier sensing nor an acknowledgment mechanism for SSs whose BRs have successfully been transmitted. Therefore, in the case of BR collision the SS will have to wait until the T16 period expires. Then it assumes a collision and retransmits its BR.

To deal with retransmissions the SS uses the *truncated binary exponential back-off (TBEB)* algorithm. According to it, in the event of failing to receive a grant the SS doubles its contention window and retransmits the BR after deferring for a number of TOs. This number is randomly chosen from the interval [0, W_1-1], where $W_1 = 2W_0$ denotes the new window size. After the i-th failure the window size is $W_i = 2^{\min\{m,i\}}W_0$. Here m implicitly defines the maximum contention window $2^m W_0$, which is specified in the Uplink Channel Descriptor (UCD) along with W_0. For each BR there is a maximum allowable number of retransmission attempts. If this limit is reached the data associated with the BR shall be dropped.

Finally, the standard also defines an optional *piggyback* procedure. According to it, an SS, after receiving a grant, may request more BW by piggybacking a collision-free BR to the data instead of contending.

B. Analytical models for Contention-based Access in WiMAX

Several researchers have looked into the contention-based access mechanism of IEEE 802.16 networks. Naturally, the proposed analytical models can be classified into two broad categories; one focuses on broadcast/group polling [3]–[14] while the other studies the contention-based CDMA BRs mechanism [19]–[24]. Interestingly, most efforts in both categories do not consider the optional piggyback process even though it can be used with any of the aforementioned contention methods.

Overall, the analysis of piggybacking has received little attention [7], [8], [16], [17] despite the fact that it can bring significant performance improvements [7], [15]. More specifically, the first modeling attempt assumes the contention-based CDMA BRs as the mandatory mechanism [16], [17]. The authors consider the queueing performance of an SS with an exhaustive queue service. To this end, they use an M/G/1 model and a Markov chain for modeling transmission periods. The authors also consider two disciplines for bandwidth allocation; transmitting one [16] or multiple [17] packets per BW grant. Although the proposed model brought some important concepts to the analysis of piggyback, some of its assumptions significantly limit its accuracy. The most important is the

TABLE I
NETWORK PARAMETERS AND NOTATION USED IN THE ANALYSIS

Network Parameters			
N	number of SSs in the network	L	number of data slots in the UL subframe
N_s	number of TOs in the UL subframe	W_0	initial contention window size ($\geq N_s$)
$2^m W_0$	maximum contention window size	W_i	contention window size in round i
M	maximum of frames in T16 period	D	maximum number of retransmissions for a BR
G	maximum number of piggybacked BRs	λ	mean rate of packet generation at each SS
$\mathrm{T_{fr}}$	frame duration	ρ_{in}	($= \lambda \mathrm{T_{fr}}$) offered load
Performance related notation			
p	BR collision probability in a TO	q	probability of receiving a BW grant in a frame
q_{M}	probability of granting BW within M successive frames	p_f	probability of not receiving BW in a contention round
$\mathrm{E}\{S\}$	expected service delay	$\mathrm{E}\{S_C\}$	expected service delay through contention
$\mathrm{E}\{S_Q\}$	expected service delay seen by queued packets	$\mathrm{E}\{W\}$	expected waiting delay
ρ	utilization ($= \lambda \mathrm{E}\{S\}$)	G	mean number of packets transmitted with piggybacking
Π_0	probability that an arriving packet finds an empty queue	P_{B}	probability that an SS transmits a BR in a frame
P_{S}	probability that a TO contains a successful BR	P_{D}	probability of dropping a BR
Th	total throughput seen by an SS	Th_{C}	throughput achieved through the contention mechanism
Th_{G}	throughput achieved through piggybacking		

assumption that the available UL BW is unlimited. In a real system the mandatory contention-based mechanism shares the same limited BW with piggybacking. Therefore the operation of one interferes with the other and vice-versa. For example, the amount of BW allocated for piggybacking influences the probability that the BS will grant BW to a contention-based BR. This in turn affects the probability that the requesting SS will issue another BR. Thus, assuming unlimited UL BW results in a less realistic modeling because it is not possible to capture the coupling between piggybacking and the contention mechanism. Moreover, the authors determine the piggyback probability based on the average queue size instead of the actual queue status. They also consider an exhaustive service, i.e. piggyback is utilized until the SS's queue is empty. This assumption deviates from a real system especially if we consider that under limited BW such a policy allows SSs using piggyback to drive contending ones to starvation. Another effort for analyzing piggybacking, this time using broadcast polling as the mandatory mechanism, has been presented in [8]. Still, this method also assumes unlimited UL BW and some sort of exhaustive service, i.e. there is no apparent limitation on how many times an SS may use piggybacking after receiving a BW grant. More importantly, the proposed model is actually semi-analytical since it requires that the request collision probability, a key performance index of broadcast polling, is determined by simulation. Finally, the authors implicitly assume a trivial queueing capability for each SS, i.e. the model considers only one of all possible arrivals during an SS's busy period. A more realistic model has been proposed by He et.al. [7]. It includes an analysis of broadcast polling, although somehow simplified (e.g. the T16 period is only one frame long). It also assumes limited UL BW as well as a more practical piggyback policy. Yet, the model bears a significant limitation; it assumes no queuing capability at the SSs and considers piggybacking only for transmitting packets that require BW allocation in successive frames. This is of paramount importance since piggybacking is mostly useful for transmitting queued packets.

In this work we put emphasis on a more realistic modeling of piggybacking in order to unveil all aspects of its performance and its synergy with the mandatory broadcast polling mechanism under both saturated and non-saturated traffic. To this end, we assume limited UL BW as well as SSs with non-trivial queueing capability. This clearly differentiates our work from other efforts studying piggybacking. Moreover, our approach also provides an improved analysis of the broadcast polling mechanism. Indeed, so far, various researchers have provided analytical models that capture the performance of broadcast/group polling only in saturated conditions [3], [5], [9] or in unsaturated conditions with the assumption of unlimited BW [4], [8], [13], [14], i.e. they only consider the contention process and not the data transmission phase. Only a set of more recent studies assumes both saturated and non-saturated traffic conditions as well as limited UL BW [6], [7], [10]–[12]. In this case one critical task is to accurately model the data transmission phase and predict the probability that the BS will not allocate a grant to a successful BR due to BW depletion. The models proposed in [6] and [10] assume that the latter probability is constant and known beforehand. In [12], the probability is calculated only under the assumption that a single BR is served in each frame. Finally, in both [7] and [11], a T16 period of one frame is assumed, i.e. only BRs received in one frame are considered for BW allocation and older BRs are dropped. This significantly affects the probability of receiving a grant. Our model moves one step forward by considering the most generic case, i.e. the BS provides grants to multiple BRs (both contending and piggybacked) in each frame while the T16 period can be greater than one frame.

III. SYSTEM MODEL AND ASSUMPTIONS

Our aim is to present a unified analytical model able to seamlessly portray the performance of broadcast polling with and without piggybacking. For this reason, we examine an IEEE 802.16 network operating in the PMP mode under either the SC or the OFDM PHY. We consider the scenario of a single BS and N SSs. The frame is structured either in the FDD or the TDD mode and its duration is $\mathrm{T_{fr}}$. The UL subframe consists of N_s TOs that SSs can use to transmit their BRs. For resolving collisions, the SSs implement the TBEB algorithm with initial window W_0, maximum window $2^m W_0$ and a maximum of D retries after which the BR is dropped. Since the optimal W_0 is not specified by the standard we reasonably assume that $W_0 \geq N_s$. Otherwise the SSs will never choose the $N_s - W_0$ remaining TOs in their first transmission attempt. An SS can piggyback a BR on data for receiving a BW grant in the next frame. However, we assume that the SS can use piggybacking in up to G consecutive frames. This is a reasonable strategy, also used in the literature [7], in order to avoid driving other SSs to starvation. The UL subframe contains L data bursts (hereafter called data slots for simplicity). In general, L is not sufficient for serving all successful BRs. Therefore the probability q of providing a grant to a BR during a specific frame is not always one. Finally, we consider a T16 period of M frames. Table I summarizes the system parameters and the notation used in the analysis. Furthermore, we adopt the following assumptions:

(A1): Each SS has a single BE data connection.

According to the IEEE 802.16 standard, BW is always requested on a connection basis but it is allocated to the SS. A1 allows us to focus on the access mechanism without going into the details that each manufacturer may implement for internal BW allocation at the SS. This assumption or a similar one, where multiple BE connections are treated as a single one, are common in the literature [4], [6], [8], [12]–[14].

(A2): Packets are generated according to a Poisson process (with rate λ) and buffered to the SS's queue. A BR is generated when a packet arrives at an empty queue or when the SS's queue is not empty and an opportunity for BR transmission exists, either through contention or piggybacking.

We follow this reasonable approach since there is no universal algorithm for creating BRs based on the incoming traffic. An alternative would be to directly model the BR arrival process. The reasonable approach in this case is to assume a memoryless process [6], [12]–[14] unless a specific algorithm for producing BRs is known. However, note that an algorithm for producing BRs would certainly consider the system state, which results in an non memoryless process. Therefore, similar to many researchers [7], [8], [11], [25], we choose to directly model the packet arrival process.

(A3): An SS's queue is infinite.

We make the assumption of an infinite queue, which is common in the literature [5], [8], [12]–[14], in order to explore the potential of piggybacking at its full extent. Note that in the reasonable case that the queue size is much greater than G then the limiting factor is actually G and the impact of the queue size is minimal.

(A4): The probability p of a BR collision in a TO is constant and independent of the SS's retransmission history.

This assumption is frequently found in the literature [5]–[7], [9]–[14] and although it may not be entirely valid we will show, when validating our analysis, that its impact on the model accuracy is minimal.

(A5): The BS allocates BW grants of one data slot that correspond to the transmission of one data packet.

Currently, there is no widely accepted bandwidth allocation mechanism for implementation at the BS. We adopt this frequently used assumption [5], [9], [11], [12], [25] in order to focus on the access mechanism without the need to examine fairness issues related to the bandwidth allocation problem. We feel that such issues are outside the scope of this work. Furthermore, note that for determining q, i.e. the probability of granting BW to a successful BR in a specific frame, the important parameter is the number of BRs that can be served in a frame. Including this assumption in our system model results in the BS being able to serve up to L BRs per frame.

(A6): Pending BRs are served by the BS using a random order service discipline.

The service discipline at the BS is important for determining q when the BS can queue a BR and serve it within $\mathrm{M}{>}1$ frames. As mentioned, most research efforts assume either unlimited UL BW and thus q=1 [3], [4], [8], [13], [14], or a predefined q [6], [10], or even limited UL BW with M=1 [9], [11]. In all these cases the service policy is trivial. Only Giambene et. al. [12] assume a Round-Robin discipline for serving BRs from different SSs. In this case the service discipline is critical because they assume that only one SS may be served during a frame. Since we allow the BS to serve multiple SSs in each frame and each BR to be served within M frames, we expect the impact of a priority based discipline to be rather limited in our case. Therefore, we adopt the simple approach of randomly choosing the BR to serve which results in a constant q that does not depend on the BR's waiting history. In general deciding on the optimal queue service discipline is a complex procedure that depends on the bandwidth allocation algorithm implemented at the BS. We believe that investigating this issue is outside the scope of this paper and plan to explore this in future work.

In the following, our analytical approach is to first examine the medium access mechanism at the SS level, i.e. analyze the process implemented by an SS (Section IV). To this end, we model the SS as an M/G/1 queueing system with vacations and the medium access procedure, including piggyback, as a Markov chain. Later, in Section V, we combine the Markov chain and well-established concepts from the analysis of IEEE 802.11 networks [18] to investigate the performance of the random access scheme at the network level.

IV. Modeling the activity of an SS

To design the Markov chain that portrays the activity of an SS we first observe that each SS can be in one of the following phases during a specific frame:

W waiting due to a back-off period
R transmitting a BR
C waiting due to a BR collision in a previous frame
F waiting for a BW grant without success after a successful BR in a previous frame
T transmitting data after receiving a BW grant as a result of a successful BR in a previous frame

Let $\mathcal{S} = \{\mathrm{W}, \mathrm{R}, \mathrm{C}, \mathrm{F}, \mathrm{T}\}$ denote the set of the aforementioned phases. Observe that each SS may contend in several rounds and each contention round consists of several frames. Furthermore, during a frame in a specific round an SS may be in any of the phases in $\mathcal{S}$. Therefore, in the proposed Markov chain we use a generic state $(i,j)^{\mathrm{X}}, X \in \mathcal{S}$ such that:

Definition 1 (State $(i,j)^{\mathrm{X}}$): The state in which the SS is in phase X during the j-th frame of the i-th contention round. For example, the interpretation of state $(i,j)^{\mathrm{F}}$ is that the SS did not receive any BW grant in the j-th frame of the

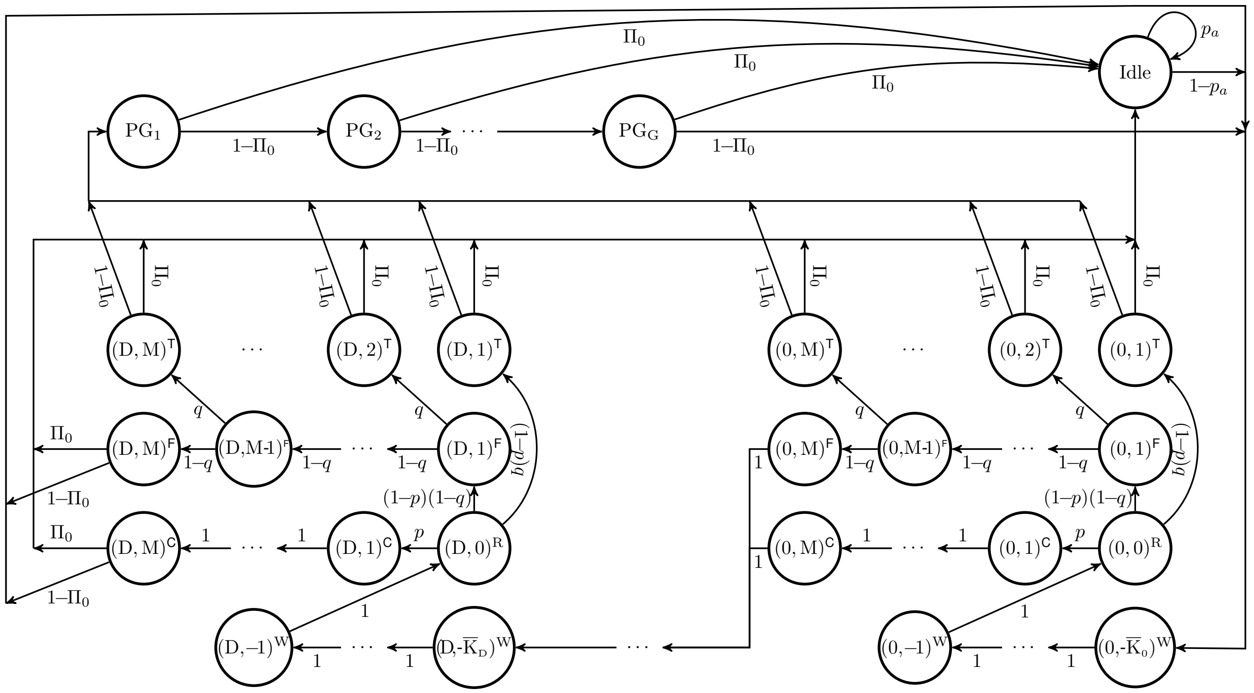


Fig. 1. The proposed Markov chain

i-th contention round although in a previous frame of the same contention round the SS transmitted successfully a BR. Furthermore, to model the piggyback mechanism as well as the case of an idle SS we also use the following states in the Markov chain:

PG$_i$ the SS transmits after receiving the i-th consecutive BW grant using the piggyback mechanism
Idle the SS waits until a new BR is produced

The proposed Markov chain is illustrated in Fig. 1. When a BR is created, the terminal chooses a value for the back-off counter in the range $[0, W_0-1]$. Therefore, the SS may wait for more than a frame before the counter expires. Let $\overline{\mathrm{K}}_0$ denote the average number of the waiting frames before the frame in which the counter expires. We model the waiting period with the sequence of states $(0,j)^{\mathrm{W}}, j = -\overline{\mathrm{K}}_0, \ldots, -1$ where the transition probability for traversing those states is one. After visiting the aforementioned states, the SS finally reaches state $(0,0)^{\mathrm{R}}$, which corresponds to the frame in which the back-off counter expires and the SS transmits the BR. This BR will be caught up in a collision with probability p, in which case the SS will wait for M frames (states $(0,j)^{\mathrm{C}}, j = 1, \ldots, \mathrm{M}$ with transition probability from state $(0, j-1)^{\mathrm{C}}$ to state $(0,j)^{\mathrm{C}}$ being one) before moving to the next contention round. In the case that the BR is successful (probability $1-p$) the SS will move from state $(0,0)^{\mathrm{R}}$ either to state $(0,1)^{\mathrm{T}}$ or $(0,1)^{\mathrm{F}}$, depending on whether BW is granted (probability q) or not (probability $1-q$). Therefore, the transition probabilities from $(0,0)^{\mathrm{R}}$ to $(0,1)^{\mathrm{T}}$ and $(0,1)^{\mathrm{F}}$ are $(1-p)q$ and $(1-p)(1-q)$ respectively. If no BW is granted (state $(0,1)^{\mathrm{F}}$) the SS will wait for a BW grant in the next frame. Consequently, the SS will move to state $(0,2)^{\mathrm{T}}$ or state $(0,2)^{\mathrm{F}}$, with probabilities q and $1-q$ respectively, depending on whether BW is granted or not. The SS may wait up to M frames to receive a BW grant. If no BW is awarded during all of these frames then the BR is considered not successful and the SS again moves to the next contention round (departure from state $(0,\mathrm{M})^{\mathrm{F}}$).

The SS repeats the same process in the following contention round. In general, there are at maximum $\mathrm{D}+1$ rounds (D retransmissions) for completing a BR transmission and receiving a BW grant. After that, the packet corresponding to the BR is dropped. The only difference between successive rounds is the contention window used by the SS. Recall that when in contention round i, an SS will randomly choose the back-off counter $backoffcnt \in [0, 2^{\min\{i,m\}}W_0 - 1]$ to indicate the number of TOs to wait before transmitting a BR. As a result, the BR will be transmitted in the $(backoffcnt{+}1)$-th consecutive TO. In general, this means that the SS will wait on average $\overline{\mathrm{K}}_i$ frames before reaching state $(i,0)^{\mathrm{R}}$ because it is possible that $backoffcnt{+}1 > N_s$. To calculate $\overline{\mathrm{K}}_i$, let us model the number of waiting frames with a discrete random variable (RV) K_i. Then, $\mathrm{J}_i = \mathrm{K}_i + 1$ is also a RV that includes the frame in which the BR is actually transmitted, i.e. the frame corresponding to state $(i,0)^{\mathrm{R}}$. Observe that the sample space of J_i is $[1, \mathrm{WC}_i]$, where $\mathrm{WC}_i = \lceil 2^{\min\{i,m\}}W_0/N_s \rceil$ (Fig. 2). Moreover, all values in $[1, \mathrm{WF}_i]$, where $\mathrm{WF}_i = \lfloor 2^{\min\{i,m\}}W_0/N_s \rfloor$ and $\mathrm{WC}_i{-}1 \leq \mathrm{WF}_i \leq \mathrm{WC}_i$, are equiprobable with probability $N_s/2^{\min\{i,m\}}W_0$. This is because for any value $x \in [1, \mathrm{WF}_i]$ there are exactly N_s values of $backoffcnt$ that will result in $\mathrm{J}_i = x$. In the case that $\mathrm{WC}_i = \mathrm{WF}_i{+}1$, i.e. when $2^{\min\{i,m\}}W_0/N_s$ is not an integer, $\mathrm{J}_i = \mathrm{WC}_i$ occurs with probability $(2^{\min\{i,m\}}W_0 - \mathrm{WF}_i N_s)/2^{\min\{i,m\}}W_0$. Consequently:

$$\begin{aligned}\overline{\mathrm{K}}_i = \overline{\mathrm{J}}_i - 1 &= \sum_{j=1}^{\mathrm{WC}_i} j\mathrm{P}\{\mathrm{J}_i = j\} - 1 \\ &= \frac{1}{2}\frac{\mathrm{WF}_i(\mathrm{WF}_i+1)N_s}{2^{\min\{i,m\}}W_0} + \mathrm{WC}_i\left(1 - \frac{\mathrm{WF}_i N_s}{2^{\min\{i,m\}}W_0}\right) - 1\end{aligned} \quad (1)$$

Let us now go back to the states $(i,j)^{\mathrm{T}}, \forall j \in [1,\mathrm{M}], i \in [0,\mathrm{D}]$, i.e. when the SS manages to successfully transmit data. The SS will now move to the idle state if no packet is waiting in the queue to be served. Let Π_0 denote this latter probability. Clearly, the SS will initiate the piggyback mechanism with probability 1–Π_0 (transition to state PG_1) and will continue to transmit data using the piggyback mechanism (states PG_2 to PG_G) if at the end of each transmission there is at least one packet waiting (probability $1-\Pi_0$). If after completing a piggyback transmission there is no waiting packet (probability Π_0) then the SS moves to the idle state. Note that the use of the piggyback mechanism is limited, therefore the SS, after G piggyback sessions, moves to the idle state with probability Π_0 (no packet is waiting) or enters contention to transmit a new BR (probability $1-\Pi_0$). To calculate Π_0, we model the SS queue as an M/G/1 queueing system with vacations [26], where packets are the customers and the service time is the total time that the SS is involved in transmitting a BR (including the backoff period) as well as the corresponding packet. Here, the vacation time is deterministic and equals $\mathrm{T_{fr}}$ because an SS with an empty queue at the start of a frame will pause serving packets arriving during this frame and will resume at the beginning of the next frame. In the aforementioned system, the probability that a departing customer leaves no customers in the system is (see Appendix A for proof):

$$\Pi_0 = (1-\rho)\frac{1-e^{-\lambda \mathrm{T_{fr}}}}{\lambda \mathrm{T_{fr}}} \tag{2}$$

where $\rho = \lambda \mathrm{E}\{S\}$, $\mathrm{E}\{S\}$ is the mean service time and $\rho_{in} = \lambda \mathrm{T_{fr}}$ the offered load. Note that (2) holds for an arbitrary departing customer [26], [27]. Thus, Π_0 and the complementary probability $1-\Pi_0$ are the transition probabilities in any case of service completion and regardless of whether the packet is served using piggyback or the contention mechanism. For the same reason, we use Π_0 in the case of a dropped packet. Recall that this happens when the corresponding BR is not served after $D{+}1$ consecutive contention rounds (states $(\mathrm{D},\mathrm{M})^{\mathrm{C}}$ and $(\mathrm{D},\mathrm{M})^{\mathrm{F}}$). From the M/G/1 model's point of view, at that point the system concludes serving the packet therefore the SS moves to the idle state with probability Π_0 or enters a new contention round to serve a new waiting packet (probability $1-\Pi_0$). Finally, when the SS is in the idle state it will remain in that state with probability $p_a = e^{-\lambda \mathrm{T_{fr}}}$, i.e. if no packet arrival occurs during the duration of a frame, or will engage in contention otherwise.

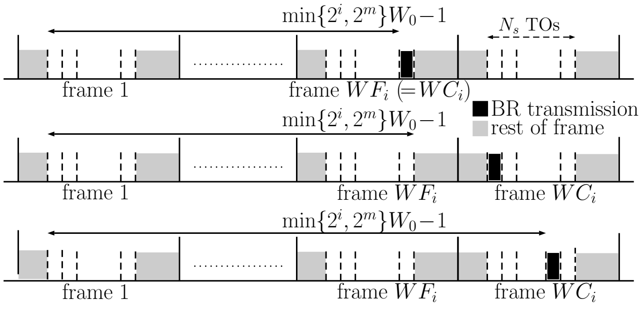


Fig. 2. Maximum number of frames in a backoff period

V. Performance Analysis of the Access Scheme

After modeling the procedure carried out by an SS we wish to use this model for analysing the performance of the access scheme at the network level. Note that the probability $b_{i,j}^{\mathrm{X}}$ of the SS being in state $(i,j)^{\mathrm{X}}$ of the proposed Markov chain can be expressed as a function of $b_{0,0}^{\mathrm{R}}$. The latter depends on three indicators (see Appendix B) that portray the performance of the access scheme, namely: i) the collision probability p, ii) the mean service time $\mathrm{E}\{S\}$ and, iii) the probability q of granting BW to a successfully transmitted contending BR. Our aim is to use the analysis of the access scheme at the network level so as to express each of the aforementioned indicators as a function of $b_{0,0}^{\mathrm{R}}$. In this way, it is possible to determine p, $\mathrm{E}\{S\}$ and q by solving a system of three non-linear equations.

A. Collision Probability

Let P_{B} denote the probability that an SS will try to transmit a BR in a specific frame. Clearly, this is the probability that the SS will transmit a BR in any contention round, i.e. the sum of the probabilities of the SS being in states $(i,0)^{\mathrm{R}}, i \in [0,\mathrm{D}]$. Therefore (for proof refer to (B.2), Appendix B):

$$P_{\mathrm{B}} = \sum_{i=0}^{\mathrm{D}} b_{i,0}^{\mathrm{R}} = \frac{1-(p_f)^{\mathrm{D}+1}}{1-p_f} b_{0,0}^{\mathrm{R}} = \tau b_{0,0}^{\mathrm{R}} \tag{3}$$

where $p_f = p + (1-p)(1-q)^{\mathrm{M}}$. Provided that an SS transmits a BR, the probability of not being involved in a collision (1–p) is the probability that none of the remaining $N{-}1$ subscribers will try to transmit in the same TO. Clearly, the number of subscribers trying to transmit in a specific TO can be modeled as a binomial RV. Therefore,

$$\begin{aligned} p &= 1-\mathrm{P}\{\text{no transmiting SS out of } N-1\} \\ &= 1-(1-\frac{P_{\mathrm{B}}}{N_s})^{N-1} = 1-(1-\frac{\tau}{N_s} b_{0,0}^{\mathrm{R}})^{N-1} \end{aligned} \tag{4}$$

where we have also assumed that it is equiprobable for an SS to select any TO.

B. Mean Service Delay

There are two alternatives for serving a packet, i.e. either through a contending or a piggybacked BR. In the case of contention, a BR will be successful and a packet will be transmitted if the SS ends up in one of the states $(i,j)^{\mathrm{T}}, \forall i \in [0,\mathrm{D}], j \in [1,\mathrm{M}]$. The probability of reaching state $(i,j)^{\mathrm{T}}$ provided that an SS creates a BR can be written (using (B.5)):

$$\begin{aligned} bb_{i,j} &= \mathrm{P}\{\text{SS in } (i,j)^{\mathrm{T}} | \text{ SS creates a BR}\} \\ &= \frac{b_{i,j}^{\mathrm{T}}}{b_{0,0}^{\mathrm{R}}} = q(1-p)(1-q)^{j-1} p_f^i \end{aligned} \tag{5}$$

Using the Markov chain we can find that the time elapsed from the start of contention until the SS reaches state $(i,j)^T$ is (expressed in frames):

$$\mathrm{E}\{S_{i,j}\} = j + i\mathrm{M} + \sum_{k=0}^{i} \overline{\mathrm{K}}_k, \quad \forall j \in [1,\mathrm{M}], i \in [0,\mathrm{D}] \tag{6}$$

The packet is dropped if the SS ends up in one of the states $(\mathrm{D},\mathrm{M})^{\mathrm{C}}$ and $(\mathrm{D},\mathrm{M})^{\mathrm{F}}$. This happens with probability

$$\begin{aligned} P_{\mathrm{D}} &= \mathrm{P}\{\text{packet dropped} \mid \text{SS creates a BR}\} \\ &= \frac{b^{\mathrm{C}}_{\mathrm{D,M}} + b^{\mathrm{F}}_{\mathrm{D,M}}}{b^{\mathrm{R}}_{0,0}} = (p_f)^{\mathrm{D}+1} \end{aligned} \tag{7}$$

while the time spent is:

$$\mathrm{E}\{S_{\mathrm{D}}\} = (\mathrm{D}+1)\mathrm{M} + \sum_{k=0}^{\mathrm{D}} \overline{\mathrm{K}}_k. \tag{8}$$

As a result, the expected service time for a packet served through the contention mechanism is:

$$\mathrm{E}\{S_{\mathrm{C}}\} = (p_f)^{\mathrm{D}+1}\mathrm{E}\{S_{\mathrm{D}}\} + \sum_{i=0}^{\mathrm{D}}\sum_{j=1}^{\mathrm{M}} bb_{i,j}\mathrm{E}\{S_{i,j}\} \tag{9}$$

For packets served through the piggyback mechanism the service delay is clearly equal to one frame, i.e. $S_{\mathrm{P}} = 1$. Note that each packet transmitted using the contention mechanism is followed on average by $\bar{G}$ packet transmissions enabled through piggybacked BRs. Therefore the overall expected service delay is

$$\mathrm{E}\{S\} = \frac{\mathrm{E}\{S_{\mathrm{C}}\} + \bar{G}\mathrm{E}\{S_{\mathrm{P}}\}}{1+\bar{G}}$$

Since in each piggyback state PG_i corresponds one data packet then:

$$\bar{G} = \sum_{i=1}^{\mathrm{G}} \mathrm{P}\{\mathrm{PG}_i | \text{SS creates a BR}\} = \sum_{i=1}^{\mathrm{G}} \frac{\mathrm{P}\{\mathrm{PG}_i\}}{b^{\mathrm{R}}_{0,0}} = Z \tag{10}$$

where we used (B.10). As a result,

$$\begin{aligned} \mathrm{E}\{S\} &= \frac{\mathrm{E}\{S_{\mathrm{C}}\} + Z}{1+Z} \\ &= \frac{(p_f)^{\mathrm{D}+1}\mathrm{E}\{S_{\mathrm{D}}\} + \sum_{i=0}^{\mathrm{D}}\sum_{j=1}^{\mathrm{M}} bb_{i,j}\mathrm{E}\{S_{i,j}\} + Z}{1+Z} \end{aligned} \tag{11}$$

C. Probability of Bandwidth Allocation

Assume that we examine a successful BR that awaits BW allocation in a specific frame and let $\mathcal{R}$ denote the total number of BRs awaiting BW allocation. Also let $\mathcal{Q}$ denote the same number excluding the examined BR, i.e. $\mathcal{R} = \mathcal{Q} + 1$. Also, let $\mathcal{P}$ denote the number of data slots to be allocated to piggybacked BRs. Recall that piggybacked BRs are served with priority over contending BRs. This policy is necessary for guaranteeing bandwidth allocation to piggybacked BRs. Therefore, $\mathcal{P}$ data slots are allocated to the SSs that are in the piggyback mode while $L - \mathcal{P}$ remaining data slots are to be allocated to $\mathcal{R}$ pending BRs. Clearly, the event of not serving some of the pending BRs occurs when $\mathcal{R} > L - \mathcal{P}$ with $\mathcal{R} - (L - \mathcal{P})$ out of the $\mathcal{R}$ BRs not receiving BW. Let I_b be the indicator RV associated with the event of not allocating BW to the examined BR. Under the assumption that all BRs are served with the same probability,

$$\mathrm{P}\{I_b = 1 | \mathcal{R} = k, \mathcal{P} = l\} = \frac{k - (L-l)}{k} \tag{12}$$

i.e., this is the probability that the examined BR is one of the $k - (L-l)$ that will not be served out of k pending BRs. Then, the probability $r = 1 - q$ of not granting BW to the examined BR is

$$\begin{aligned} r &= \mathrm{P}\{I_b = 1 | \mathcal{R}, \mathcal{P}\}\mathrm{P}\{\mathcal{R} > L - \mathcal{P}\} \\ &= \sum_i \mathrm{P}\{I_b = 1 | \mathcal{R}, \mathcal{P} = i\}\mathrm{P}\{\mathcal{R} > L - i\}\mathrm{P}\{\mathcal{P} = i\} \\ &= \sum_i \mathrm{P}\{I_b = 1 | \mathcal{R}, \mathcal{P} = i\}\mathrm{P}\{\mathcal{Q} + 1 > L - i\}\mathrm{P}\{\mathcal{P} = i\} \\ &= \sum_i \sum_{j > L-i-1} \Big[\mathrm{P}\{I_b = 1 | \mathcal{R} = j+1, \mathcal{P} = i\} \\ &\qquad \mathrm{P}\{\mathcal{Q} = j | \mathcal{P} = i\}\mathrm{P}\{\mathcal{P} = i\}\Big] \end{aligned} \tag{13}$$

In order to determine r (or equivalently q) we should determine both $\mathrm{P}\{\mathcal{P} = i\}$ and $\mathrm{P}\{\mathcal{Q} = j | \mathcal{P} = i\}$. It is possible to model the allocation of data slots to the piggyback mechanism using a set of L independent Bernoulli RVs. In this case, we can model $\mathcal{P}$ as a binomial RV, therefore

$$\mathrm{P}\{\mathcal{P} = i\} = \binom{L}{i}\mathrm{P}^i_G(1 - P_{\mathrm{G}})^{L-i} \tag{14}$$

where $P_{\mathrm{G}} = \mathrm{E}\{\mathcal{P}\}/L$ is the probability that a data slot is allocated to piggybacking. Observe that in any given frame an SS has either one pending or a piggybacked BR. Therefore, if $\mathcal{P} = i$ SSs are in piggyback mode and one SS has already produced the examined BR then there are $N - i - 1$ SSs that may have a pending BR. As a result, we can model $\mathcal{Q}$ as a sum of $N - i - 1$ Bernoulli RVs. Therefore the conditional pmf of $\mathcal{Q}$ given $\mathcal{P}$ is Binomial, i.e.

$$\mathrm{P}\{\mathcal{Q} = j | \mathcal{P} = i\} = \binom{N-i-1}{j}\mathrm{P}^j_A(1 - P_{\mathrm{A}})^{N-i-1-j} \tag{15}$$

where $P_{\mathrm{A}} = \mathrm{E}\{\mathcal{R}\}/(N-i)$ is the probability of an SS that is not in piggyback mode to have a pending BR. By combining (13) with (12),(14) and (15) we conclude that

$$\begin{aligned} r = \sum_{i=0}^{L}\sum_{j=L-i}^{N-i-1} \Bigg[& \frac{j+1-(L-i)}{j+1}\binom{N-i-1}{j}P^j_{\mathrm{A}} \\ & (1 - P_{\mathrm{A}})^{N-i-1-j}\binom{L}{i}P^i_{\mathrm{G}}(1 - P_{\mathrm{G}})^{L-i}\Bigg] \end{aligned} \tag{16}$$

Note that it is possible to derive q using (16) as long as we determine both P_{G} and P_{A} or equivalently $\mathrm{E}\{\mathcal{P}\}$ and $\mathrm{E}\{\mathcal{R}\}$. In Appendix C we prove that:

$$\mathrm{E}\{\mathcal{R}\} = N_s P_{\mathrm{S}} \sum_{i=0}^{\mathrm{M}-1}(1-q)^i \tag{17}$$

and

$$\mathrm{E}\{\mathcal{P}\} = \mathrm{E}\{\mathcal{R}\}q\sum_{i=0}^{\mathrm{G}-1}(1 - \Pi_0)^{i+1} \tag{18}$$

where P_{S} is the probability of a successful BR transmission in a TO. In order to calculate P_{S} observe that a successful BR is one not involved in a collision. In other words, a TO contains a successful BR if only one SS transmits in the slot, therefore,

$$P_{\mathrm{S}}=N\frac{P_{\mathrm{B}}}{N_s}(1-\frac{P_{\mathrm{B}}}{N_s})^{N-1}=N\frac{\tau b^{\mathrm{T}}_{0,0}}{N_s}(1-\frac{\tau b^{\mathrm{T}}_{0,0}}{N_s})^{N-1} \qquad (19)$$

where P_{B} is the probability that an SS transmits a BR in a frame and is given by (3).

D. Determining System Performance

Up to this point we have managed to come up with three equations that involve only three variables. More specifically, observe that equations (4), (11) and (16) form a system of three equations where only the performance indicators p, $\mathrm{E}\{S\}$ and q are unknown. Therefore, it is possible to numerically solve this system and determine these indicators as well as other performance metrics that up to now we have managed to express as a function of these indicators such P_{S}, $\bar{G}$, P_{D}, etc. It is also possible to extend our analysis in order to explore more performance aspects of the access scheme. One such interesting aspect is the overall achieved throughput. Recall that the probability of a TO containing a successful BR is P_{S} and can be calculated using (19). At the same time, a BW grant may be allocated to a successful BR in one of the M consecutive frames of the T16 period with probability q per frame. Therefore, the probability that a successful BR is actually served in a contention round is $q_{\mathrm{M}} = 1-(1-q)^{\mathrm{M}}$. Consequently, the throughput achieved through contention, expressed in packets per frame, is:

$$Th_{\mathrm{C}} = [1-(1-q)^{\mathrm{M}}]P_{\mathrm{S}}N_s = q_{\mathrm{M}}P_{\mathrm{S}}N_s \qquad (20)$$

To calculate the throughput achieved by means of the piggyback mechanism bear in mind that each attempt from an SS to transmit a contending BR (which occurs with probability $b^{\mathrm{R}}_{0,0}$) finally succeeds with probability $1-P_{\mathrm{D}}$. Furthermore, each packet sent after a successful BR is followed on average by $\bar{G}=Z$ packets sent using piggybacked BRs. As a result, the throughput of the piggyback mechanism is:

$$Th_{\mathrm{PG}} = ZN(1-P_{\mathrm{D}})b^{\mathrm{R}}_{0,0} \qquad (21)$$

where Z, P_{D} and $b^{\mathrm{R}}_{0,0}$ are given by (10), (7) and (B.12) respectively. The overall achieved throughput is obviously $Th = Th_{\mathrm{C}} + Th_{\mathrm{PG}}$.

Another very interesting performance metric is the time that an arriving packet has to wait before being served. As we mentioned earlier, we model an SS's queue as an M/G/1 system with vacations. Therefore it is reasonable to consider the Pollaczek-Khinchine formula as a starting point. However, in our system the service delay depends on whether an arriving packet enters the queue or not. In Appendix D we approximate the expected waiting delay as:

$$\mathrm{E}\{W\} \simeq \frac{\lambda\mathrm{E}\{S^2\}}{2(1-\lambda\mathrm{E}\{S_Q\})} + \frac{\mathrm{T_{fr}}}{2} \qquad (22)$$

where $\mathrm{E}\{S_Q\}$ is the expected service time seen by a queued packet and is given by (D.3), while $\mathrm{E}\{S^2\}$ is the second order moment of the expected service delay and can be calculated using (D.2). Note that (22) can be seen as a modified version of the Pollaczek-Khinchine formula that results if we replace $\mathrm{E}\{S\}$ with $\mathrm{E}\{S_Q\}$ in the original formula.

VI. Results & Model Validation

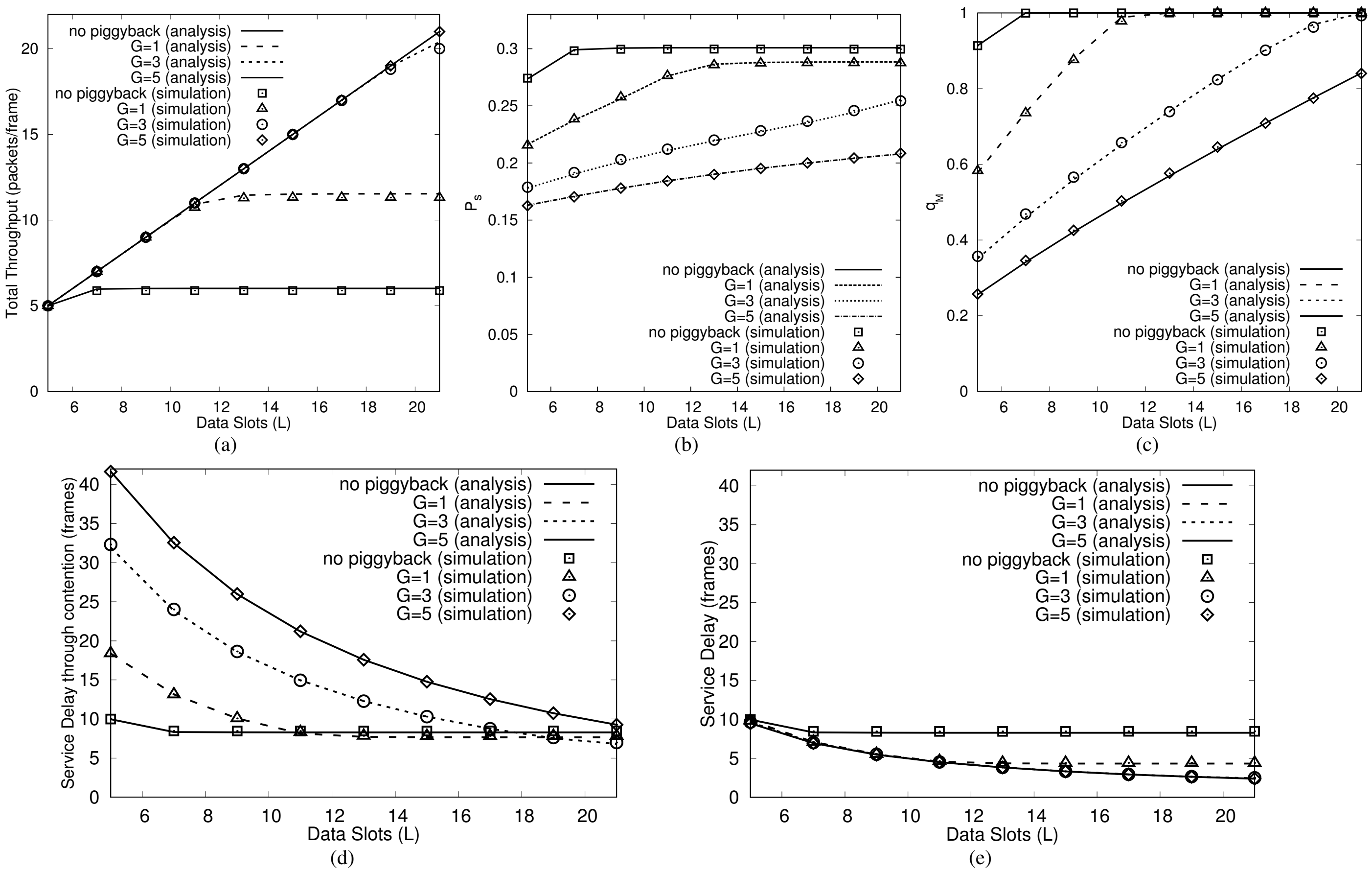
In order to investigate the impact of the piggyback mechanism on the IEEE 802.16 contention-based access scheme as well as to verify the proposed analytical model we conducted several simulation experiments. To this end, we used a custom event-driven simulator, written in C++, that implements broadcast polling with the piggyback mechanism as specified in the IEEE 802.16 standard [1]. Table II presents the parameters used in our simulations. Since we focus on exploring the features of piggybacking, we refrain from investigating the impact of parameters associated with broadcast polling such as W_0, N_s, etc. The role of such parameters has been extensively investigated in the literature [6], [12]. Therefore we choose typical default values based on this literature. In any case, the interested reader can refer to Appendix E where we present a more extensive set of experiments, including ones exploring the impact of the aforementioned parameters. Note that we use $N_s = 20 < W_0$. In this case the probability that an SS transmits a BR is not the same for all TOs. However, we make this choice to illustrate the accuracy of our model even in such a challenging scenario. Finally, to provide a more generic analysis we use the frame duration as the time unit. For example, throughput related metrics are expressed on a per frame basis. Similarly, the unit for delay related metrics is the frame duration.

In the first experiment we focus on capturing the full extent of the benefits associated with the piggyback mode. In particular, we investigate the performance of the piggyback-enabled random access method versus the number of data slots L in saturation conditions, i.e. $\rho_{in} = \lambda\mathrm{T_{fr}} = 1$. Fig. 3a illustrates the overall throughput (attained either through contention or piggyback) for different values of G. Observe that when piggyback is disabled (i.e. $\mathrm{G}=0$) the saturation throughput is achieved for $L=7=\lceil 0.33N_s \rceil$. This result has previously been found in [25] where the authors prove that the optimal data slots to TOs ratio is $\frac{L}{N_s} = \frac{\ln 2}{2} \approx 1/3$. Note that without piggyback the saturation throughput is achieved when the probability of successful BR transmission (P_{S}) is maximized (Fig. 3b). When $L < 7$ the throughput decreases since the available data slots are not sufficient for satisfying the offered load. This is confirmed by Fig. 3c where $q_{\mathrm{M}} < 1$ when $L < 7$, i.e. it is possible for a successful BR not to receive a BW grant even if it waits for M frames.

On the other hand, when piggyback is enabled ($\mathrm{G} > 0$) the number of data slots required for maximizing throughput is clearly greater than $0.33N_s$. This is reasonable if we bear in mind that each BW grant allocated to a successful BR is followed on average by the transmission of $\bar{G}$ packets through piggybacking. Thus, apart from the BW allocated to successful BRs, BW is also required for the piggyback mechanism. A rough estimation for the number of data slots required for

TABLE II
Parameters used in Simulations

<table>
<tr><td colspan="5">Range of parameters</td></tr>
<tr><td>$N \in [2, 50]$</td><td>$L \in [5, 21]$</td><td colspan="2">$\lambda \in [0.05, 2]$ packets/frame
$\rho_{in} \in [0.05, 1]$ Erlang</td><td>$\mathrm{G} \in [0, 5]$</td></tr>
<tr><td colspan="5">Default values</td></tr>
<tr><td>$N_s = 20$</td><td>$W_0 = 32$</td><td>$m = 5$</td><td>$\mathrm{M} = 6$</td><td>$\mathrm{D} = 5$</td></tr>
</table>



(a) (b) (c)

(d) (e)

Fig. 3. Performance vs number of data slots (L) under saturated conditions ($\rho_{in} = 1$): (a) Total Throughput (Th) (b) P_{S}, (c) q, (d) Service delay through contention ($\mathrm{E}\{S_{\mathrm{C}}\}$) and (e) Service delay ($\mathrm{E}\{S\}$)

maximizing throughput in saturation conditions is given by

$$L'(\mathrm{G}) = \lceil (\mathrm{G}+1)\mathrm{P}_{\mathrm{S}}^{\max} N_s \rceil \tag{23}$$

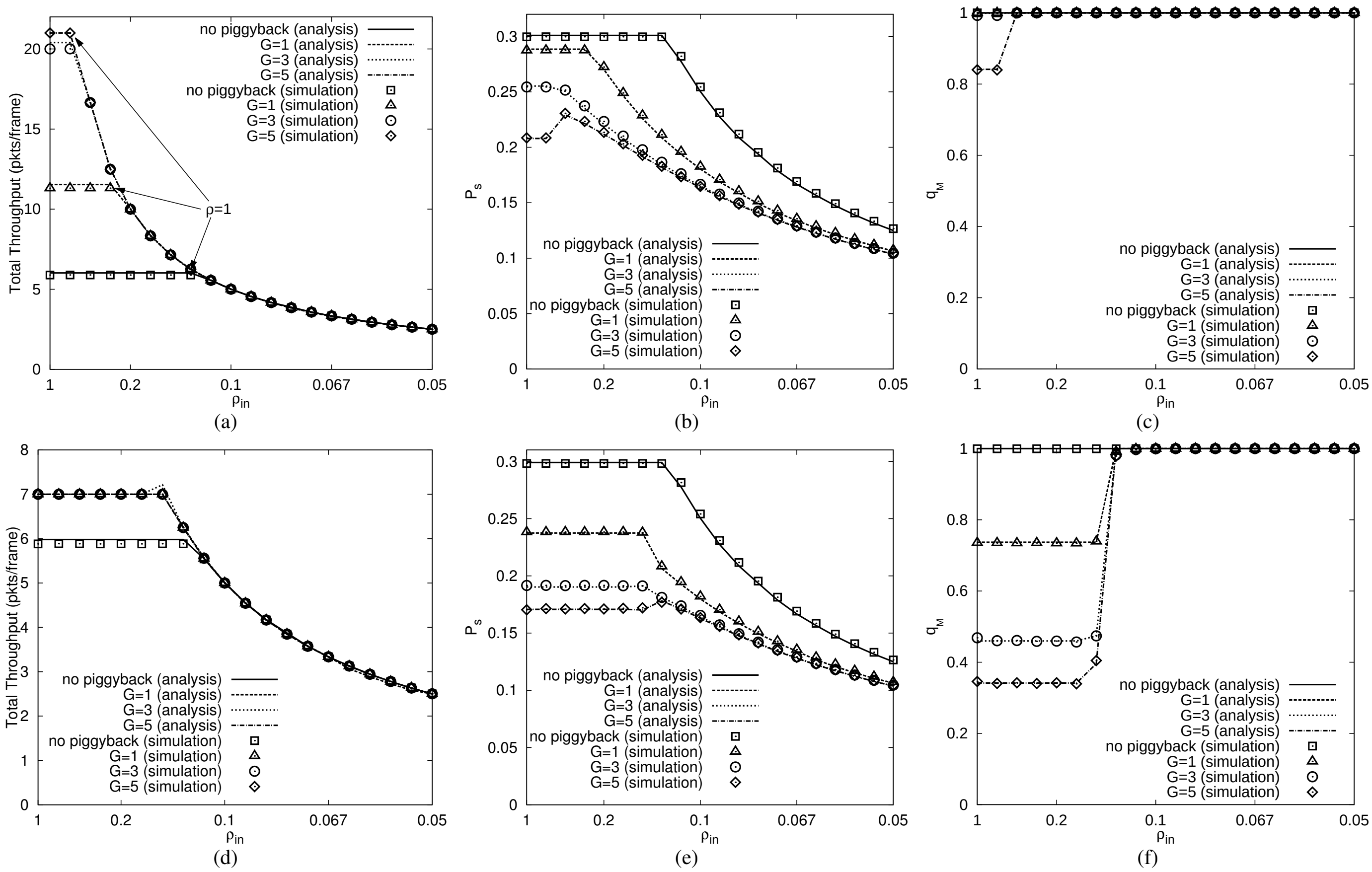
where $\mathrm{P}_{\mathrm{S}}^{\max}$ is the maximum probability of successful BR transmission. We also use G instead of $\bar{G}$ because in saturation $\bar{G} = \mathrm{G}$. Observe that $\mathrm{P}_{\mathrm{S}}^{\max} N_s$ is actually the average number of successful BRs while $\mathrm{G}\mathrm{P}_{\mathrm{S}}^{\max} N_s$ is the BW required for packets sent using piggybacked BRs. Similarly, the maximum throughput (in pkts/frame) can be approximated as

$$Th^{\mathrm{MAX}} = \min\{L, L'(\mathrm{G})\} = \min\{L, (\mathrm{G}+1)\mathrm{P}_{\mathrm{S}}^{\max} N_s\} \tag{24}$$

In general, when $L < L'(\mathrm{G})$ throughput is upper bounded by L and no further improvement is possible regardless of whether we increase G because it is not possible to exceed L pkts/frame. For example, as illustrated in Fig. 3a, when $\mathrm{G}=1$ the throughput is upper bounded by L for every value $L < L'(1) = 12$. Therefore, there is no point in using $\mathrm{G}>1$ when $L < 12$ because in all cases the throughput would not exceed L pkts/frame (note that performance is identical for all $\mathrm{G} > 0$ when $L < 12$). To explain this result from another point of view recall that in each frame $\mathrm{E}\{\mathcal{P}\}$ data slots are occupied by the piggyback mechanism and only $L - \mathrm{E}\{\mathcal{P}\}$ are available for successful contending BRs. As G increases $L - \mathrm{E}\{\mathcal{P}\}$ is cut back. Consequently, not only it is less probable that a BR will receive a grant (Fig. 3c) but also fewer SSs are involved in transmitting new contending BRs (observe the reduction of P_{S} in Fig. 3b). The latter is because more SSs are entangled in long waiting periods in order to receive a BW grant and therefore they become idle. Overall, fewer packets are transmitted through contention. This reduces the piggyback throughput since piggybacking is only possible after the transmission of a packet through contention. In order to break this negative feedback cycle and improve throughput we need to increase both L and G. In this way, we avoid reducing $L - \mathrm{E}\{\mathcal{P}\}$ excessively and therefore allow q_{M} to increase. In other words, to obtain a throughput gain we should increase L beyond $L'(1) = 12$ and then use $\mathrm{G} > 1$.

Going back to the case that $L \geq L'(\mathrm{G})$, it is easy to show (by combining (C.2) and (C.1)) that in saturation $L - \mathrm{E}\{\mathcal{P}\} > P_{\mathrm{S}} N_s$, i.e. there is always enough bandwidth to serve all successful BRs (i.e. $q_{\mathrm{M}} = 1$), therefore it is possible to maximize throughput. This is the case for example when $\mathrm{G} = 1$ and $L \geq L'(1) = 12$ as can be seen in Fig. 3a-3c. However, note that the upper bound of throughput does not increase linearly with G. For example, for $G = 1$ the upper bound is ~ 11.3 pkts/frame while for $G = 3$ it is only ~ 20 pkts/frame. According to (24) the upper bound depends on G, $\mathrm{P}_{\mathrm{S}}^{\max}$ and N_s. Interestingly, for the reasons explained previously, $\mathrm{P}_{\mathrm{S}}^{\max}$ itself depends on G, i.e. for a certain L value $\mathrm{P}_{\mathrm{S}}^{\max}$ decreases as G increases (Fig. 3b). This explains the non-linear dependency of maximum throughput and G.

In Fig. 3d we present the service delay for packets transmitted through the contention mechanism ($\mathrm{E}\{S_C\}$) while Fig. 3e illustrates the service delay for all transmitted packets ($\mathrm{E}\{S\}$). Evidently, $\mathrm{E}\{S_C\}$ is higher when $\mathrm{G} > 0$ compared to the



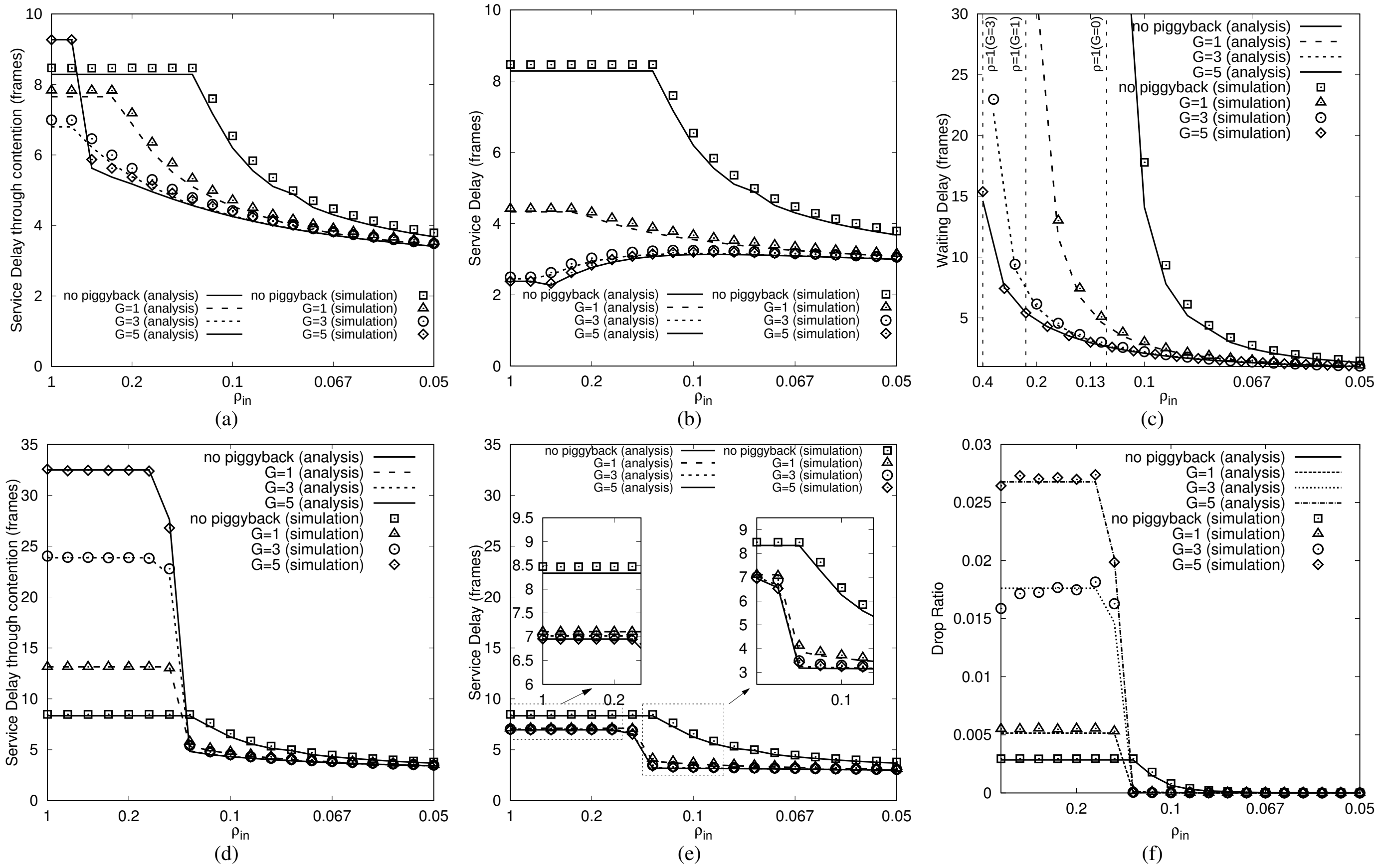
(a) (b) (c)

(d) (e) (f)

Fig. 4. Throughput related metrics vs ρ_{in} for $L = 21$ ((a)-(c)) and $L = 7$ ((d)-(f)): (a),(d) Total Throughput (Th) (b),(e) P_{S}, and (c),(f) q

case that piggyback is disabled unless a considerable amount of bandwidth is available. This is a direct indication of the more severe competition for available bandwidth when the latter is not in abundance. Indeed, as mentioned previously, when G > 0 only $L - \mathrm{E}\{\mathcal{P}\}$ data slots are available for successfully contending BRs. Hence, it is more difficult to receive a BW grant in a specific frame (Fig. 3c). As a result, on average more frames are required for receiving a BW grant, thus increasing service delay. The situation is reversed when bandwidth is not an issue (see Fig. 3d, $L = 21$). In such cases the piggyback mechanism takes full advantage of bandwidth availability and significantly reduces the delay. We provide more insight into this performance aspect in the next experiment. Another important observation is that $\mathrm{E}\{S_C\}$ increases dramatically for smaller values of L as well as for greater values of G since in both cases the L–$\mathrm{E}\{\mathcal{P}\}$ available data slots are reduced dramatically. Nevertheless, increasing G has a positive impact on the overall service delay $\mathrm{E}\{S\}$. The reason is that more packets are transmitted with minimum delay using the piggyback mechanism. In accordance with our observations so far, obtaining delay gains requires increasing both G and L at the same time. Nonetheless, using piggyback to reduce delay involves also a downside. The delay jitter for two successive packets increases when one packet is sent using piggybacking while the other uses a contending BR.

In the next experiment we examine the system performance with respect to the offered load. For this, we use two cases with different bandwidth availability; $L = 7$ and $L = 21$. As discussed in the previous experiment, the first value is the minimum that allows throughput maximization when piggyback is disabled. At the same time, $L = 21$ is adequate for illustrating the performance differentiation for various values of G. Fig. 4a and 4d present the total throughput for $L = 21$ and $L = 7$ respectively. Reasonably, in both cases, when the offered load is relatively low there is no performance differentiation regardless of whether we use piggyback (with any value of G) or not. Since the offered load is low the probability of successive arrivals that could trigger the piggyback mechanism, i.e. those with small temporal separation, is low. The benefits of piggybacking appear when the offered load increases. However, in the case of $L = 7$ the improvement is limited and the same saturation throughput is achieved for all values of G. On the contrary, in the case of $L = 21$ a greater value of G results in a higher maximum throughput. Note that saturation, i.e. $\rho = 1$, manifests itself at an increasing level of offered load for higher values of G (e.g. $\rho_{in} \approx 0.25$ for G$=1$, $\rho_{in} \approx 0.5$ for G$=3$). In other words, piggybacking manages to improve system stability, i.e. $\rho < 1$ for a wider range of offered load. This is in contrast to the $L = 7$ case where for all values of G saturation appears when $\rho_{in} \approx 0.14$. The basic reason for witnessing the contrasting performance features in the two cases is BW availability. More specifically, when $L = 7$ saturation is clearly the result of limited BW availability and the discussion in the previous experiment can again explain this performance. In fact, all supporting evidence can be found in Fig. 4e and 4f; q_{M} decreases with G while P_{S} is smaller



(a) (b) (c)

(d) (e) (f)

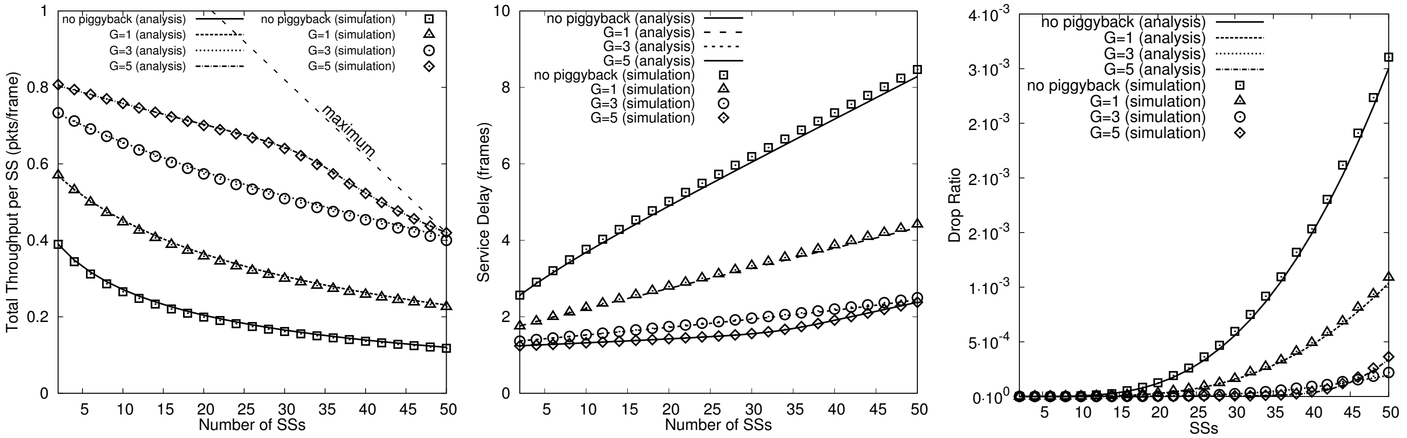
Fig. 5. Delay related metrics vs ρ_{in} for $L=21$ ((a)-(c)) and $L=7$ ((d)-(f)): (a),(d) Service Delay through contention (E$\{S_{\mathrm{C}}\}$) (b),(e) Service Delay (E$\{S\}$), (c) Waiting Delay (E$\{W\}$), (f) Drop Ratio (P_{D})

compared to the case with more available bandwidth (i.e. $L=21$, Fig. 4b). On the other hand, the $L=21$ case portrays a more "healthy" operation mode. System saturation comes as the result of the limited capabilities of the IEEE 802.16 contention access scheme. Indeed, with the exception of G$\,=5$ (our simulation revealed that $L\approx 35$ is required to achieve maximum throughput in this case), the system allocates BW to all successfully contending BRs (Fig. 4c) and the maximum throughput is determined by P_{S} which is a characteristic of the contention mechanism.

The smoother operation in the $L = 21$ case is also evident in the delay-related performance metrics (Fig. 5a-5c). Piggybacking not only manages to reduce the overall service delay but it also significantly reduces the delay for packets transmitted through contention. Since a significant amount of traffic is forwarded using piggybacked BRs, congestion in the basic random access mechanism is alleviated. Furthermore, every contending BR receives a BW grant in the next frame ($q_{\mathrm{M}}=1$). The immediate consequence is that an SS needs to wait for fewer frames, thus the reduced E$\{S_C\}$. Offloading traffic through piggyback has also a positive influence on the waiting delay in the queue. Using higher G values drastically reduces it (Fig. 5c). Note again that the system stability is improved. i.e. $\rho<1$ for a wider range of the offered load. Concluding, when sufficient bandwidth exists it is always beneficial to use piggyback and the benefits are more important for higher G values. This is not the case however when the bandwidth is rather limited (e.g. $L=7$). Although increasing G reduces the overall delay (Fig. 5e) the gains are limited. More importantly, the service delay provided by the contention method increases abruptly when the offered load is high. This results in a significant delay jitter compared to the piggyback mechanism. Finally, another downside is that there is a non-negligible probability of dropping a BR (Fig. 5f).

So far, the discussion in the two experiments, besides the advantages of piggybacking, highlights the need for a careful consideration when choosing G. The optimal value should always be chosen after considering the available bandwidth. Choosing a high value for G may have a significant impact on the smooth operation of the basic contention mechanism.

Finally, we examine performance against the number of SSs that participate in the network when $L = 21$ and $\rho_{in} = 1$ (Fig. 6). Again, it is evident that the piggyback mechanism can provide a more efficient operation. In particular, although the throughput seen by each SS (Fig. 6a) reasonably decreases as more SSs participate in the network, the use of piggyback always allows an SS to sustain a higher throughput. Similar gains are witnessed with respect to service delay (Fig. 6b) or the probability of dropping a BR (Fig. 6c). Note that the gains obtained by using G$\,=5$ instead of G$\,=3$ are limited when the number of SSs is close to 50. This is because in this case the system's bandwidth is not adequate for supporting such an extensive use of piggybacking (as mentioned previously, when G$\,=5$ performance is maximized for $L\approx 35$). This is another confirmation that one should consider the available BW when choosing the optimal G.



(a) (b) (c)

Fig. 6. Performance vs number of SSs (N) for $L=21$ under saturated conditions ($\rho_{in} = 1$): (a) Total Throughput per SS (Th/N) (b) Service Delay ($\mathrm{E}\{S\}$), (c) Drop Ratio (P_{D})

VII. Conclusion

In this work, we proposed and validated through extensive simulations an analysis of broadcast polling when the piggyback mechanism is employed by SSs. Unlike other analytical efforts, our model covers not only the contention phase but also the bandwidth allocation one. Moreover, it places emphasis on using a more accurate modeling of the system in general and especially for those of its aspects that are directly associated with piggybacking. This, on one hand, enables us to analyze piggybacking and its synergy with broadcast polling under more realistic conditions. On the other hand, it also facilitates a more accurate model for plain broadcast polling. The detailed performance evaluation of piggybacking reveals that it can potentially bring significant performance improvements. However, it also brings to light the associated trade-offs regarding the delay jitter as well as the need for a careful allocation of UL BW between contending and piggybacked BRs.

Appendix A
Probability of an Empty Queue seen by a Departing customer

In a generic M/G/1 system with vacations the probability that an arbitrary departing customer leaves the system empty is given by [26]:

$$\Pi_0 = (1-\rho)/F'(1) \tag{A.1}$$

where $F(z) = \sum_{j=1}^{\infty} f_j z^j$ is the probability generating function for the number of customers awaiting service when the server returns from vacation to find at least one customer waiting. In other words,

$$\begin{aligned} f_j &= \mathrm{P}\{j \text{ customers waiting after vacation } | j \neq 0\} \\ &= \frac{\mathrm{P}\{j \text{ customers arrive during last vacation}\}}{\mathrm{P}\{j \neq 0\}} \end{aligned}$$

Note that the number of customers waiting at the end of vacation is the number of customers that arrived during the last vacation interval because at the beginning of that last interval the number of waiting customers was zero (otherwise there would be no vacation). In our case the vacation duration is deterministic with value $\mathrm{T_{fr}}$. Therefore,

$$f_j = \frac{\mathrm{P}(j;\lambda\mathrm{T_{fr}})}{1-e^{\lambda\mathrm{T_{fr}}}}$$

where $\mathrm{P}(j;\lambda\mathrm{T_{fr}})$ is the probability of j arrivals during a frame. Note that $F'(1)=\mathrm{E}(j)$ and

$$\mathrm{E}(j) = \frac{\sum_{j=1}^{\infty} j\mathrm{P}(j;\lambda\mathrm{T_{fr}})}{1-e^{\lambda\mathrm{T_{fr}}}}$$

Note that the sum in the previous equation is equal to the mean of Poisson distribution, therefore

$$F'(1) = \frac{\lambda\mathrm{T_{fr}}}{1-e^{-\lambda\mathrm{T_{fr}}}} \tag{A.2}$$

which in combination with (A.1) results in (2).

Appendix B
Solving the Markov Chain

Regarding the states corresponding to the transmission of a BR, i.e. $(i,0)^{\mathrm{R}}, \forall i \in [0,D]$, recall that the SS moves to contention round $i+1$ and transmits again a BR if it fails to transmit data in round i. This happens because either the SS did not receive a BW grant in M consecutive frames or its BR is caught up in a collision. Therefore the corresponding probability is $p_f = p + (1-p)(1-q)^{\mathrm{M}}$. Indeed, using the proposed Markov chain (Fig. 1) we can show that $b^R_{i+1,0} = p_f b^R_{i,0}$. By recursion we establish that:

$$b^R_{i,0} = p_f^i b^R_{0,0}, \forall i \in [1,\mathrm{D}] \tag{B.1}$$

and therefore:

$$\sum_{i=0}^{D} b^{\mathrm{R}}_{i,0} = \sum_{i=0}^{\mathrm{D}} (p_f)^i b^{\mathrm{R}}_{0,0} = \overbrace{\frac{1-(p_f)^{\mathrm{D}+1}}{1-p_f}}^{\tau} b^{\mathrm{R}}_{0,0} = \tau b^{\mathrm{R}}_{0,0} \tag{B.2}$$

For the states traversed during back-off $b^{\mathrm{W}}_{i,j} = b^{\mathrm{R}}_{i,0}, \forall j \in [-\overline{\mathrm{K}}_i, -1]$. Using (B.1):

$$\sum_{i=0}^{\mathrm{D}} \sum_{j=-\overline{\mathrm{K}}_i}^{-1} b^{\mathrm{W}}_{i,j} = \sum_{i=1}^{\mathrm{D}} \overline{\mathrm{K}}_i b^{\mathrm{R}}_{i,0} = \overbrace{\sum_{i=1}^{\mathrm{D}} (p_f)^i \overline{\mathrm{K}}_i}^{\Omega} b^{\mathrm{R}}_{0,0} = \Omega b^{\mathrm{R}}_{0,0} \tag{B.3}$$

In the case of a collision in round $i \in [0, \mathrm{D}]$ the SS will go over states $(i, j)^{\mathrm{C}}, j \in [1, \mathrm{M}]$. Note that $b^{\mathrm{C}}_{i,j} = b^{\mathrm{C}}_{i,k}, \forall j, k \in [1, \mathrm{M}]$. The probability of collision is p therefore $b^{\mathrm{C}}_{i,1} = pb^{\mathrm{R}}_{i,0}$. With the help of (B.1) we find that:

$$\sum_{i=0}^{\mathrm{D}}\sum_{j=1}^{\mathrm{M}} b^{\mathrm{C}}_{i,j} = p\mathrm{M}\sum_{i=0}^{\mathrm{D}}(p_f)^i b^{\mathrm{R}}_{0,0} = p\mathrm{M}\tau b^{\mathrm{R}}_{0,0} \tag{B.4}$$

In the case of a successful BR the possible states are $(i, j)^{\mathrm{T}}$ and $(i, j)^{\mathrm{F}}, \forall j \in [1, \mathrm{M}], i \in [0, \mathrm{D}]$ and the corresponding probabilities result from the Markov chain as:

$$\begin{aligned} b^{\mathrm{T}}_{i,j} &= (1-p)q(1-q)^{j-1} b^{\mathrm{R}}_{i,0}, \quad \forall j \in [1, \mathrm{M}], i \in [0, \mathrm{D}] \\ b^{\mathrm{F}}_{i,j} &= (1-p)(1-q)^{j} b^{\mathrm{R}}_{i,0}, \quad\quad \forall j \in [1, \mathrm{M}], i \in [0, \mathrm{D}] \end{aligned} \tag{B.5}$$

Combining these equations with (B.1), we show that:

$$\begin{aligned} \sum_{i=0}^{\mathrm{D}}\sum_{j=1}^{\mathrm{M}} b^{\mathrm{T}}_{i,j} &= \sum_{i=0}^{\mathrm{D}}\sum_{j=1}^{\mathrm{M}} q(1-p)(1-q)^{j-1} p_f^i b^{\mathrm{R}}_{0,0} \\ &= (1-p_f)\tau b^{\mathrm{R}}_{0,0} \end{aligned} \tag{B.6}$$

$$\begin{aligned} \sum_{i=0}^{\mathrm{D}}\sum_{j=1}^{\mathrm{M}} b^{\mathrm{F}}_{i,j} &= \sum_{i=0}^{\mathrm{D}}\sum_{j=1}^{\mathrm{M}} (1-p)(1-q)^{j} p_f^i b^{\mathrm{R}}_{0,0} \\ &= \frac{1-q}{q}(1-p_f)\tau b^{\mathrm{R}}_{0,0} \end{aligned} \tag{B.7}$$

Regarding the piggyback states, note that:

$$\begin{aligned} \mathrm{P}\{\mathrm{PG}_1\} &= \sum_{i=0}^{\mathrm{D}}\sum_{j=1}^{\mathrm{M}} (1-\Pi_0) b^{\mathrm{T}}_{i,j} \\ &= (1-\Pi_0)(1-p_f)\tau b^{\mathrm{R}}_{0,0} \end{aligned} \tag{B.8}$$

$$\mathrm{P}\{\mathrm{PG}_i\} = (1-\Pi_0)\mathrm{P}\{\mathrm{PG}_{i-1}\}, \forall i \in [2, \mathrm{G}] \tag{B.9}$$

therefore:

$$\begin{aligned} \sum_{i=1}^{\mathrm{G}} \mathrm{P}\{\mathrm{PG}_i\} &= \mathrm{P}\{\mathrm{PG}_1\}\sum_{i=1}^{\mathrm{G}}(1-\Pi_0)^{i-1} \\ &= \overbrace{(1-\Pi_0)(1-p_f)\tau\frac{1-(1-\Pi_0)^{\mathrm{G}}}{\Pi_0}}^{Z} b^{\mathrm{R}}_{0,0} \end{aligned} \tag{B.10}$$

The probability of being in the idle state can be calculated according to (B.11). Finally, using the normalization condition for the Markov chain, we compute $b^{\mathrm{R}}_{0,0}$ in (B.12).

Appendix C
Derivation of E{$\mathcal{P}$} and E{$\mathcal{R}$}

Let us first focus on $\mathrm{E}\{\mathcal{R}\}$ and assume that we examine the most recent frame (frame 0) from a sequence of M frames with indexes from $-(\mathrm{M}–1)$ to 0. Furthermore, let $\mathcal{R}_{-i}$ denote the number of new successful BRs through contention in frame $-i$ and $\mathcal{R}'_{-i}$ the subset of those BRs that are not served until frame 0 although waiting for i frames. Then, $\mathcal{R}$ can be expressed as the sum of new successful BRs in frame 0 as well as BRs from previous frames that are yet to be served although waiting for up to $\mathrm{M}-1$ frames, i.e., $\mathcal{R} = \mathcal{R}_0 + \mathcal{R}'_{-1} + \ldots + \mathcal{R}'_{-(\mathrm{M}-1)}$. Note that $\mathcal{R}_0$ is equivalent to the number of TOs in frame 0 that contain a successful BR. Since SSs act independently and choose a TO randomly, $\mathcal{R}_0$ can be modelled as a binomial RV, i.e., $\mathcal{R}_0 \sim B(N_s, \mathrm{P}_S)$, where $B(\cdot)$ denotes the binomial distribution, N_s is the number of TOs and P_S is the probability of a successful BR in a TO and given in (19). Observe that, in analogy, every $\mathcal{R}_{-i}$ is also binomially distributed with the same parameters. Regarding $\mathcal{R}'_{-1}$, recall that in frame -1 there are $\mathcal{R}_{-1}$ new successful BRs. Each of those BRs will not be served with probability $1-q$ and will now wait for BW allocation in frame 0. The decision for serving or not one of the $\mathcal{R}_{-1}$ BRs can be modelled with a Bernoulli RV. Since the decisions for all BRs are independent, the number of non-served BRs $\mathcal{R}'_{-1}$ is a binomial RV with parameters $\mathcal{R}_{-1}$ and $1-q$, i.e., $\mathcal{R}'_{-1} \sim B(\mathcal{R}_{-1}, 1–q)$. Since $\mathcal{R}_{-1}$ is also a binomial RV with parameters N_s and P_S it is well known that $\mathcal{R}'_{-1} \sim B(N_s, P_{\mathrm{S}}(1–q))$, i.e., $\mathcal{R}'_{-1}$ follows a binomial distribution with parameters N_s and $P_{\mathrm{S}}(1-q)$. Following a similar reasoning we conclude that:

$$\mathcal{R}'_{-i} \sim B(N_s, P_{\mathrm{S}}(1-q)^i), \forall i \in [1, \mathrm{M}-1]$$

because successful BRs in frame $-i$ reach frame 0 if and only if they do not receive a BW grant for i consecutive frames. As a result, the expected number of BRs awaiting BW allocation is

$$\begin{aligned} \mathrm{E}\{\mathcal{R}\} &= \mathrm{E}\{\mathcal{R}_0\} + \mathrm{E}\{\mathcal{R}'_{-1}\} + \cdots + \mathrm{E}\{\mathcal{R}'_{-(\mathrm{M}–1)}\} \\ &= N_s P_{\mathrm{S}} + N_s P_{\mathrm{S}}(1-q) + \cdots + N_s P_{\mathrm{S}}(1-q)^{\mathrm{M}–1} \end{aligned} \tag{C.1}$$

which results in (17).

Following a similar approach, we can express $\mathcal{P}$, i.e., the number of data slots allocated to piggypacked BRs, as

$$\begin{aligned} \mathrm{P}_{\mathrm{Idle}} &= p_a \mathrm{P}_{\mathrm{Idle}} + \Pi_0 \sum_{i=0}^{\mathrm{D}}\sum_{j=1}^{\mathrm{M}} b^{\mathrm{T}}_{i,j} + \Pi_0 \sum_{k=1}^{\mathrm{G}} \mathrm{P}\{\mathrm{PG}_i\} + \Pi_0 (b^{\mathrm{C}}_{D,M} + b^{\mathrm{F}}_{D,M}) \overset{(B.1),(B.6),(B.10)}{=} p_a \mathrm{P}_{\mathrm{Idle}} + \Pi_0(1-p_f)\tau b^{\mathrm{R}}_{0,0} \\ &+ [1-(1-\Pi_0)^{\mathrm{G}}](1-\Pi_0)(1-p_f)\tau b^{\mathrm{R}}_{0,0} + \Pi_0 (p_f)^{\mathrm{D+1}} b^{\mathrm{R}}_{0,0} = \frac{\overbrace{\Pi_0 + [1-(1-\Pi_0)^{\mathrm{G}}](1-\Pi_0)[1-(p_f)^{\mathrm{D+1}}]}^{\Phi}}{1-p_a} b^{\mathrm{R}}_{0,0} \end{aligned} \tag{B.11}$$

$$\begin{aligned} 1 &= \sum_{i=1}^{\mathrm{D}}\sum_{j=-\overline{\mathrm{K}}_i}^{-1} b^{\mathrm{W}}_{i,j} + \sum_{i=0}^{\mathrm{D}} b^{\mathrm{R}}_{i,0} + \sum_{i=0}^{\mathrm{D}}\sum_{j=1}^{\mathrm{M}} b^{\mathrm{C}}_{i,j} + \sum_{i=0}^{\mathrm{D}}\sum_{j=1}^{\mathrm{M}} b^{\mathrm{T}}_{i,j} + \sum_{i=0}^{\mathrm{D}}\sum_{j=1}^{\mathrm{M}} b^{\mathrm{F}}_{i,j} + \sum_{i=1}^{\mathrm{G}} \mathrm{P}\{\mathrm{PG}_i\} + \mathrm{P}_{\mathrm{Idle}} \\ &\overset{(B.1)-(B.11)}{=} \left\{\tau + \Omega + p\mathrm{M}\tau + (1-p_f)\tau + \frac{1-q}{q}(1-p_f)\tau + Z + \Phi\right\} b^{\mathrm{R}}_{0,0} \Rightarrow b^{\mathrm{R}}_{0,0} = \frac{1}{\tau\left(1 + p\mathrm{M} + \frac{1-p_f}{q}\right) + \Omega + Z + \Phi} \end{aligned} \tag{B.12}$$

$\mathcal{P} = \mathcal{P}_0 + \mathcal{P}'_{-1} + \ldots + \mathcal{P}'_{-(\mathrm{G}-1)}$, where $\mathcal{P}_{-i},\ i \in [0, \mathrm{G}-1]$ is the number of SSs that start piggyback in frame $-i$ and $\mathcal{P}'_{-i}$ is the number of SSs that start piggyback in frame $-i$ but are still in piggyback mode during frame 0. Observe that $\mathcal{P}_0$ is the number of SSs that had a pending BR (i.e. a successfully contending BR that awaits BW allocation) in frame -1, received BW allocation (probability q) and after completing the data transmission there was at least one packet in their queue (probability $1 - \Pi_0$). Since the number of pending BRs in a frame is $\mathcal{R}$ it is clear that $\mathcal{P}_0 \sim B(\mathcal{R}, q(1-\Pi_0))$. Therefore $\mathrm{E}[\mathcal{P}_0] = \mathrm{E}[\mathcal{R}]q(1-\Pi_0)$. Similarly, $\mathcal{P}_{-1}$ is the number of SSs that start piggyback in frame -1, therefore $\mathcal{P}_{-1} \sim B(\mathcal{R}, q(1-\Pi_0))$. Only a subset of these SSs will still be in piggyback in frame 0, i.e. those that find a new queued packet after transmitting the previous one. Therefore, $\mathcal{P}'_{-1} \sim B(\mathcal{P}_{-1}, 1-\Pi_0)$ or equivalently $\mathcal{P}'_{-1} \sim B(\mathcal{R}, q(1-\Pi_0)^2)$. As a result, $\mathrm{E}[\mathcal{P}'_{-1}] = \mathrm{E}[\mathcal{R}]q(1-\Pi_0)^2$. By generalizing $\mathrm{E}[\mathcal{P}'_{-i}] = \mathrm{E}[\mathcal{R}]q(1-\Pi_0)^{i+1}$, therefore

$$\begin{aligned}\mathrm{E}\{\mathcal{P}\} &= \mathrm{E}\{\mathcal{P}_0\} + \mathrm{E}\{\mathcal{P}'_{-1}\} + \cdots + \mathrm{E}\{\mathcal{P}'_{-(\mathrm{G}-1)}\} \\ &= \mathrm{E}\{\mathcal{R}\}q\big[(1-\Pi_0)+(1-\Pi_0)^2+\cdots+(1-\Pi_0)^{\mathrm{G}}\big]\end{aligned} \quad (\mathrm{C.2})$$

which results in (18).

Appendix D
Derivation of Expected Waiting Delay

To prove the approximation formula in (22) let us start with the mean value analysis used to prove the Pollaczek-Khinchin formula [27]. In an M/G/1 system without vacations:

$$\mathrm{E}\{W_{\mathrm{noV}}\} = \rho\mathrm{E}\{S_{\mathrm{R}}\} + N_Q\mathrm{E}\{S_{\mathrm{Q}}\} = \frac{\lambda\mathrm{E}\{S^2\}}{2(1-\lambda\mathrm{E}\{S_{\mathrm{Q}}\})} \quad (\mathrm{D.1})$$

where $\mathrm{E}\{S_{\mathrm{R}}\}$ is the remaining service time of the customer in service seen by an arriving customer, N_Q is the expected number of customers in queue and $\mathrm{E}\{S_{\mathrm{Q}}\}$ the expected service delay for customers in the queue. We have also used $N_Q = \lambda\mathrm{E}\{W_{\mathrm{noV}}\}$ from Little's law and $\mathrm{E}\{S_{\mathrm{R}}\} = \lambda\mathrm{E}\{S^2\}/2\mathrm{E}\{S\}$ [27]. In a typical M/G/1 system $\mathrm{E}\{S_{\mathrm{Q}}\} = \mathrm{E}\{S\}$ and the Pollaczek-Khinchin formula follows from (D.1). When server vacations of fixed size T_{fr} are considered it can be proved that $\mathrm{E}\{W\} = \mathrm{E}\{W_{\mathrm{noV}}\} + \mathrm{T}_{\mathrm{fr}}/2$ [27] and in combination with (D.1) we receive (22) where $\mathrm{E}\{S^2\}$ can be approximated using (11) as

$$\mathrm{E}\{S^2\} \approx \frac{(p_f)^{\mathrm{D}+1}(\mathrm{E}\{S_{\mathrm{D}}\})^2 + \sum_{i=0}^{\mathrm{D}}\sum_{j=1}^{\mathrm{M}} bb_{i,j}(\mathrm{E}\{S_{i,j}\})^2 + Z}{1+Z} \quad (\mathrm{D.2})$$

Unlike a typical M/G/1 system, in our case the expected service delay for customers in the queue is not the same as the overall expected service delay, i.e., $\mathrm{E}\{S_{\mathrm{Q}}\} \neq \mathrm{E}\{S\}$. This is because only queued packets can enjoy the low delay of piggybacking. To determine $\mathrm{E}\{S_{\mathrm{Q}}\}$ recall that an arriving packet enters the queue with probability 1–Π_0. In such a case, the new packet finds the system either busy (with probability ρ) or idle (with probability $1 - \Pi_0 - \rho$), i.e. there are other packets waiting but the server is still in vacation. Therefore

$$\mathrm{E}\{S_{\mathrm{Q}}\} = \frac{\rho}{1-\Pi_0}\mathrm{E}\{S_B\} + \frac{1-\Pi_0-\rho}{1-\Pi_0}\mathrm{E}\{S_I\} \quad (\mathrm{D.3})$$

where $\mathrm{E}\{S_B\}$and $\mathrm{E}\{S_I\}$denote the service delay experienced by the arriving packet in each of the aforementioned cases. In the case that the packet finds the system non empty but idle it is clear that the packet is not the first one waiting for service. This means that with high probability it will be served through piggybacking therefore $\mathrm{E}\{S_I\} \simeq 1$ unless $\mathrm{G} = 0$, i.e. piggyback is disabled, in which case $\mathrm{E}\{S_I\} \simeq \mathrm{E}\{S_{\mathrm{C}}\}$. Regarding $\mathrm{E}\{S_B\}$, we can rewrite the overall service delay

$$\mathrm{E}\{S\} = \Pi_0\mathrm{E}\{S_{\mathrm{C}}\} + (1-\Pi_0-\rho)\mathrm{E}\{S_I\} + \rho\mathrm{E}\{S_B\} \quad (\mathrm{D.4})$$

where the first component corresponds to an arriving packet that finds the system empty (probability Π_0) in which case it will be served through contention. It is possible to determine $\mathrm{E}\{S_B\}$ using (9) and (D.4).

Appendix E
Impact of Various Parameters

A. Experiment E-1: Performance vs load when $N_s = W_0$

In this experiment we investigate the accuracy of the model when $N_s = W_0$. This is a typical setting used in the related literature. The relation between N_s and W_0 puts under test the accuracy of analytical models. This is because it directly affects the validity of the assumptions used in calculating the probability of a successful BR, e.g. the assumption that the probability that a TO will contain a successful BR is the same for all TOs in a frame. Table III summarizes the parameters used in this experiment.

TABLE III
Parameters used in experiment E-1

Range of parameters			
$\lambda \in [0.08, 1]$ packets/frame $\rho_{in} \in [0.08, 1]$ Erlang			$\mathrm{G} \in [0, 5]$
Default values			
$N = 50$	$N_s = 32$	$W_0 = 32$	$L = 21$
$m = 5$		$\mathrm{M} = 6$	$\mathrm{D} = 5$

TABLE IV
Parameters used in experiment E-2

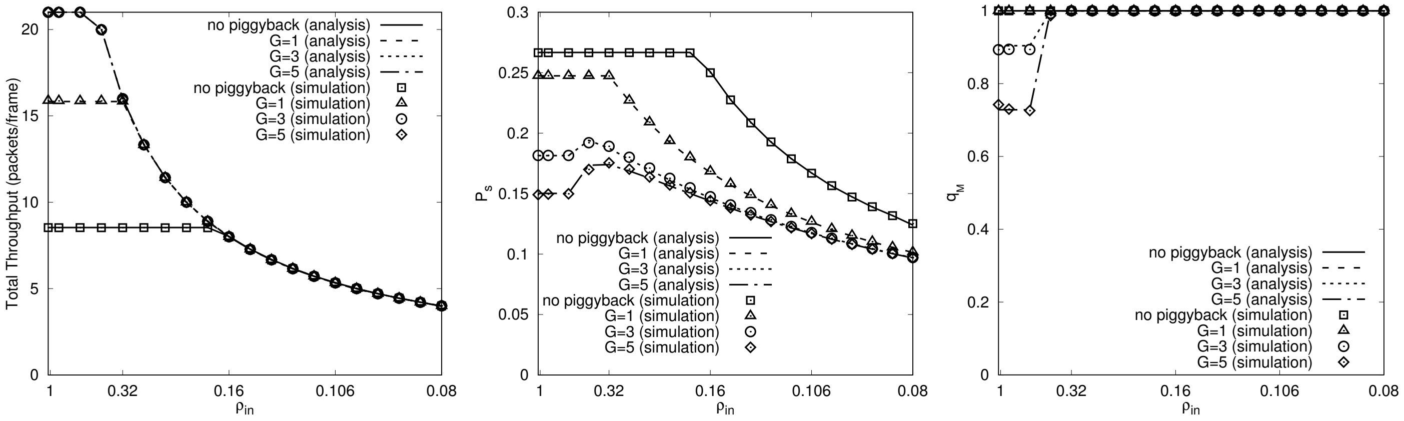
Range of parameters		
$\lambda \in [0.04, 0.8]$ packets/frame $\rho_{in} \in [0.04, 0.8]$ Erlang	$\mathrm{G} \in [0, 5]$	$W_0 \in [16, 64]$
Default values		
$N = 50$	$N_s = 16$	$L = 7$
$m = 5$	$\mathrm{M} = 6$	$\mathrm{D} = 5$

TABLE V
Parameters used in experiment E-3

Range of parameters		
$L \in [5, 21]$	$\mathrm{G} \in [0, 5]$	$\mathrm{M} \in [1, 6]$
Default values		
$\rho_{in} = 1$	$N = 50$	$N_s = 20$
$m = 5$	$W_0 = 32$	$\mathrm{D} = 5$

TABLE VI
Parameters used in experiment E-4

Range of parameters		
$L \in [5, 21]$	$\mathrm{G} \in [0, 5]$	$\mathrm{D} \in [1, 5]$
Default values		
$\rho_{in} = 1$	$N = 50$	$N_s = 20$
$m = 5$	$W_0 = 32$	$\mathrm{M} = 6$

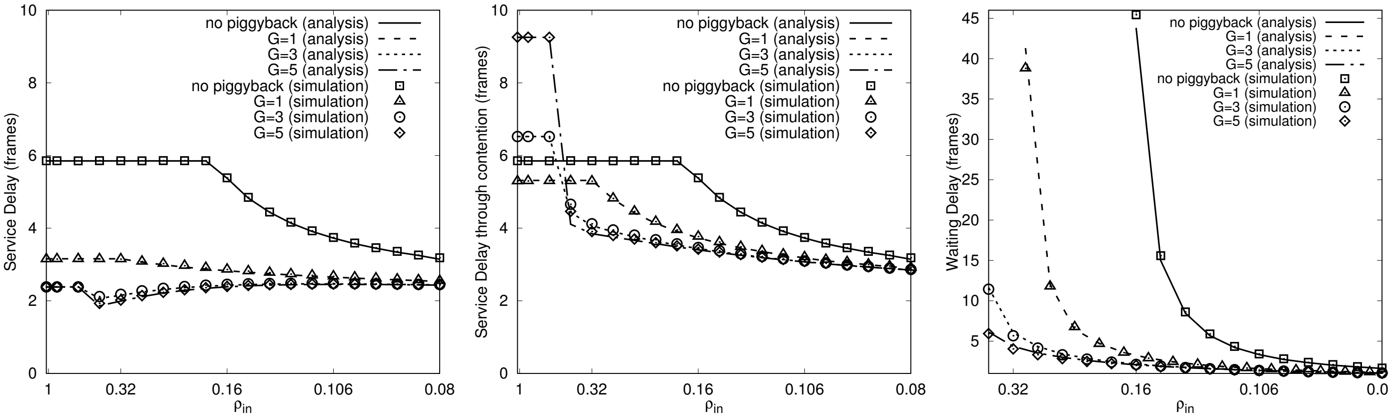


(a)

(b)

(c)

Fig. 7. Throughput related metrics vs ρ_{in} for $L = 21$ ($N_s = W_o = 32$, $m = D = 5$, $M = 6$): (a) Total Throughput (Th) (b) P_s and (c) q



(a)

(b)

(c)

Fig. 8. Delay related metrics vs ρ_{in} for $L = 21$ ($N_s = W_o = 32$, $m = D = 5$, $M = 6$): (a) Service Delay ($E\{S\}$) (b) Service Delay through contention ($E\{S_c\}$) and (c) Waiting Delay ($E\{W\}$)

B. Experiment E-2: Performance vs load for various W_0

In this experiment we investigate the impact of a growing initial contention window W_0. This investigation reveals the extent at which the TBEB algorithm affects the performance of broadcast polling and in turn the synergy of the latter with piggybacking. Table IV summarizes the parameters used in this experiment.

C. Experiment E-3: Performance vs data slots for various values of M

In this experiment we investigate performance vs the available bandwidth for different values of the T16 period. T16 is important for serving BRs when the available BW is limited. The impact of T16 becomes more important in the case of piggybacking because BW is also allocated to piggybacked BRs therefore a smaller portion of BW is available to the contention mechanism. Therefore, choosing a suitable T16 period is critical for contending BRs. Table V summarizes the parameters used in this experiment.

D. Experiment E-4: Performance vs data slots for various values of D

In this experiment we investigate the impact of the maximum number of retransmissions for a BR (D) on the system performance. Table VI summarizes the parameters used in this experiment.

References

[1] "IEEE standard for local and metropolitan area networks part 16: Air interface for broadband wireless access systems," *IEEE Std 802.16-2009 (Revision of IEEE Std 802.16-2004)*, pp. 1–2080, May 2009.

[2] C.So-In, R.Jain, and A.K.Tamimi, "Scheduling in IEEE 802.16e mobile wimax networks: key issues and a survey," *IEEE Journal on Selected Areas in Communications*, vol. 27, no. 2, pp. 156–171, February 2009.

[3] A. Vinel, Y. Zhang, M. Lott, and A. Tiurlikov, "Performance analysis of the random access in ieee 802.16," in *2005 IEEE 16th International Symposium on Personal, Indoor and Mobile Radio Communications*, vol. 3. IEEE, 2005, pp. 1596–1600.

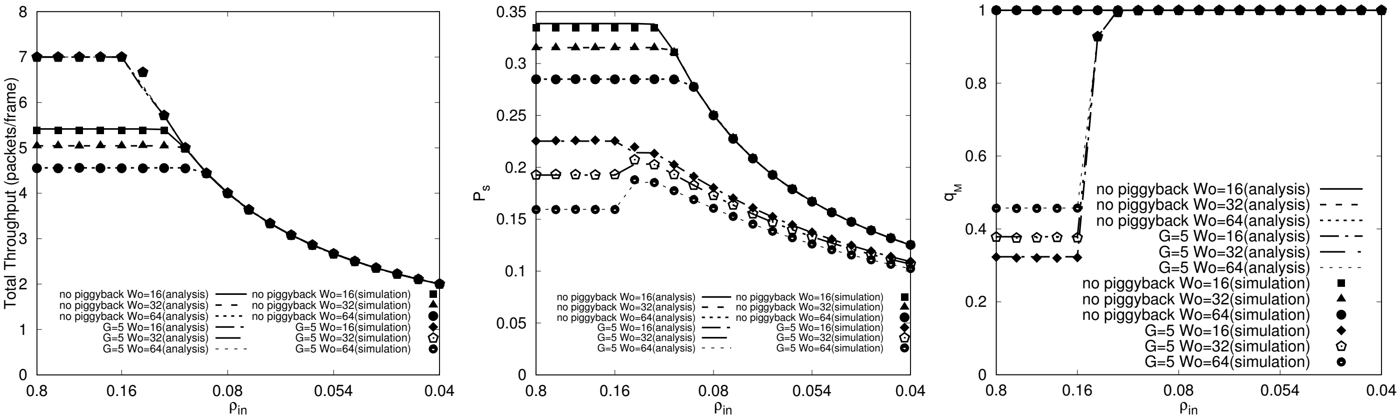
[4] A. Vinel, Y. Zhang, Q. Ni, and A. Lyakhov, "Efficient request mechanism usage in ieee 802.16," in *IEEE Globecom 2006*. IEEE, 2006, pp. 1–5.

[5] J. He, K. Guild, K. Yang, and H.-H. Chen, "Modeling contention based bandwidth request scheme for ieee 802.16 networks," *Communications Letters, IEEE*, vol. 11, no. 8, pp. 689–700, August 2007.

[6] Y. Fallah, F. Agharebparast, M. Minhas, H. Alnuweiri, and V. Leung, "Analytical modeling of contention-based bandwidth request mechanism in ieee 802.16 wireless networks," *Vehicular Technology, IEEE Transactions on*, vol. 57, no. 5, pp. 3094–3107, Sept 2008.

[7] J. He, K. Yang, K. Guild, and H.-H. Chen, "On bandwidth request mechanism with piggyback in fixed ieee 802.16 networks," *Wireless Communications, IEEE Transactions on*, vol. 7, no. 12, pp. 5238–5243, December 2008.

[8] L.-W. Chen and Y.-C. Tseng, "Design and analysis of contention-based request schemes for best-effort traffics in ieee 802.16 networks," *Communications Letters, IEEE*, vol. 12, no. 8, pp. 602–604, Aug 2008.

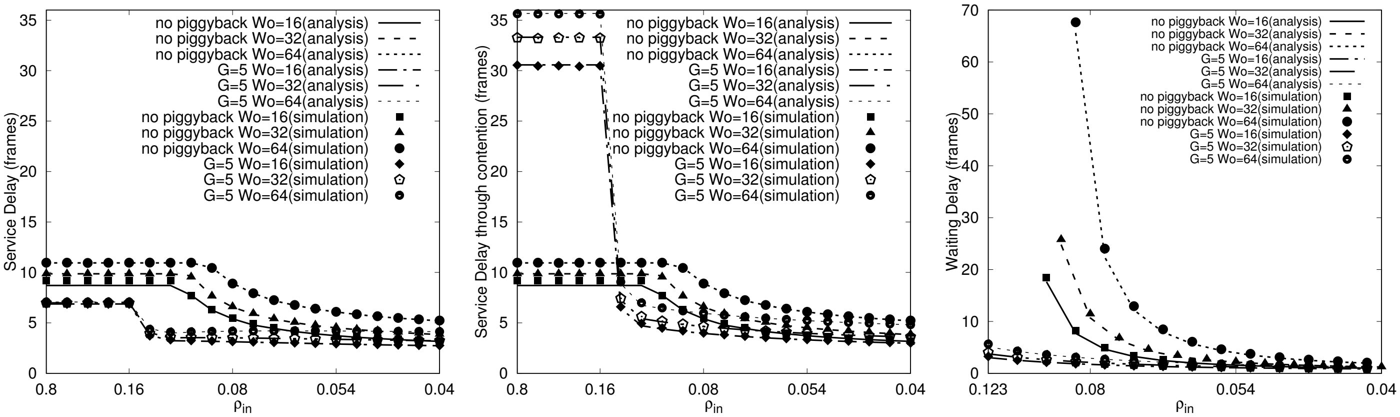


(a)

(b)

(c)

Fig. 9. Throughput related metrics vs ρ_{in} for $L = 7$ ($N_s = 16$, $m = D = 5$, $M = 6$): (a) Total Throughput (Th) (b) P_s and (c) q



(a)

(b)

(c)

Fig. 10. Delay related metrics vs ρ_{in} for $L = 7$ ($N_s = 16$, $m = D = 5$, $M = 6$): (a) Service Delay ($E\{S\}$) (b) Service Delay through contention ($E\{S_c\}$) and (c) Waiting Delay ($E\{W\}$)

[9] H. L. Vu, S. Chan, and L. L. Andrew, "Performance analysis of best-effort service in saturated ieee 802.16 networks," *Vehicular Technology, IEEE Transactions on*, vol. 59, no. 1, pp. 460–472, 2010.

[10] D. Chuck, K.-Y. Chen, and J. M. Chang, "A comprehensive analysis of bandwidth request mechanisms in ieee 802.16 networks," *Vehicular Technology, IEEE Transactions on*, vol. 59, no. 4, pp. 2046–2056, 2010.

[11] J. Liu, S. Chan, and H. L. Vu, "Performance modeling of broadcast polling in ieee 802.16 networks with finite-buffered subscriber stations," *Wireless Communications, IEEE Transactions on*, vol. 11, no. 12, pp. 4514–4523, 2012.

[12] G. Giambene and S. Hadzic-Puzovic, "Nonsaturated performance analysis for wimax broadcast polling access," *Vehicular Technology, IEEE Transactions on*, vol. 62, no. 1, pp. 306–325, 2013.

[13] Q. Ni and L. Hu, "An unsaturated model for request mechanisms in wimax," *Communications Letters, IEEE*, vol. 14, no. 1, pp. 45–47, January 2010.

[14] Q. Ni, L. Hu, A. Vinel, Y. Xiao, and M. Hadjinicolaou, "Performance analysis of contention based bandwidth request mechanisms in wimax networks," *Systems Journal, IEEE*, vol. 4, no. 4, pp. 477–486, Dec 2010.

[15] C. Cicconetti, A. Erta, L. Lenzini, and E. Mingozzi, "Performance evaluation of the ieee 802.16 mac for qos support," *Mobile Computing, IEEE Transactions on*, vol. 6, no. 1, pp. 26–38, 2007.

[16] J. B. Seo, H. W. Lee, and C. H. Cho, "Performance of ieee802.16 random access protocol - steady state queuing analysis," in *IEEE Globecom 2006*, Nov 2006, pp. 1–6.

[17] H. W. Lee and J. B. Seo, "Queueing performance of ieee 802.16 random access protocol with bulk transmissions," in *2007 IEEE International Conference on Communications*, June 2007, pp. 5963–5968.

[18] G. Bianchi, "Performance analysis of the ieee 802.11 distributed coordination function," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 3, pp. 535–547, March 2000.

[19] Y.-J. Choi, S. Park, and S. Bahk, "Multichannel random access in ofdma wireless networks," *Selected Areas in Communications, IEEE Journal on*, vol. 24, no. 3, pp. 603–613, 2006.

[20] D. Staehle, R. Pries, A. Vinel, and A. Mäder, "Performance evaluation and parameterization of the ieee 802.16 contention-based cdma bandwidth request mechanism for the ofdma physical layer," in *ACM MSWiM*, 2009, pp. 374–383.

[21] J.-B. Seo and V. Leung, "Design and analysis of backoff algorithms for random access channels in umts-lte and ieee 802.16 systems," *Vehicular Technology, IEEE Transactions on*, vol. 60, no. 8, pp. 3975–3989, 2011.

[22] J.-B. Seo and V. C. Leung, "Queuing performance of multichannel s-aloha systems with correlated arrivals," *IEEE Transactions on Vehicular Technology*, vol. 60, no. 9, pp. 4575–4586, 2011.

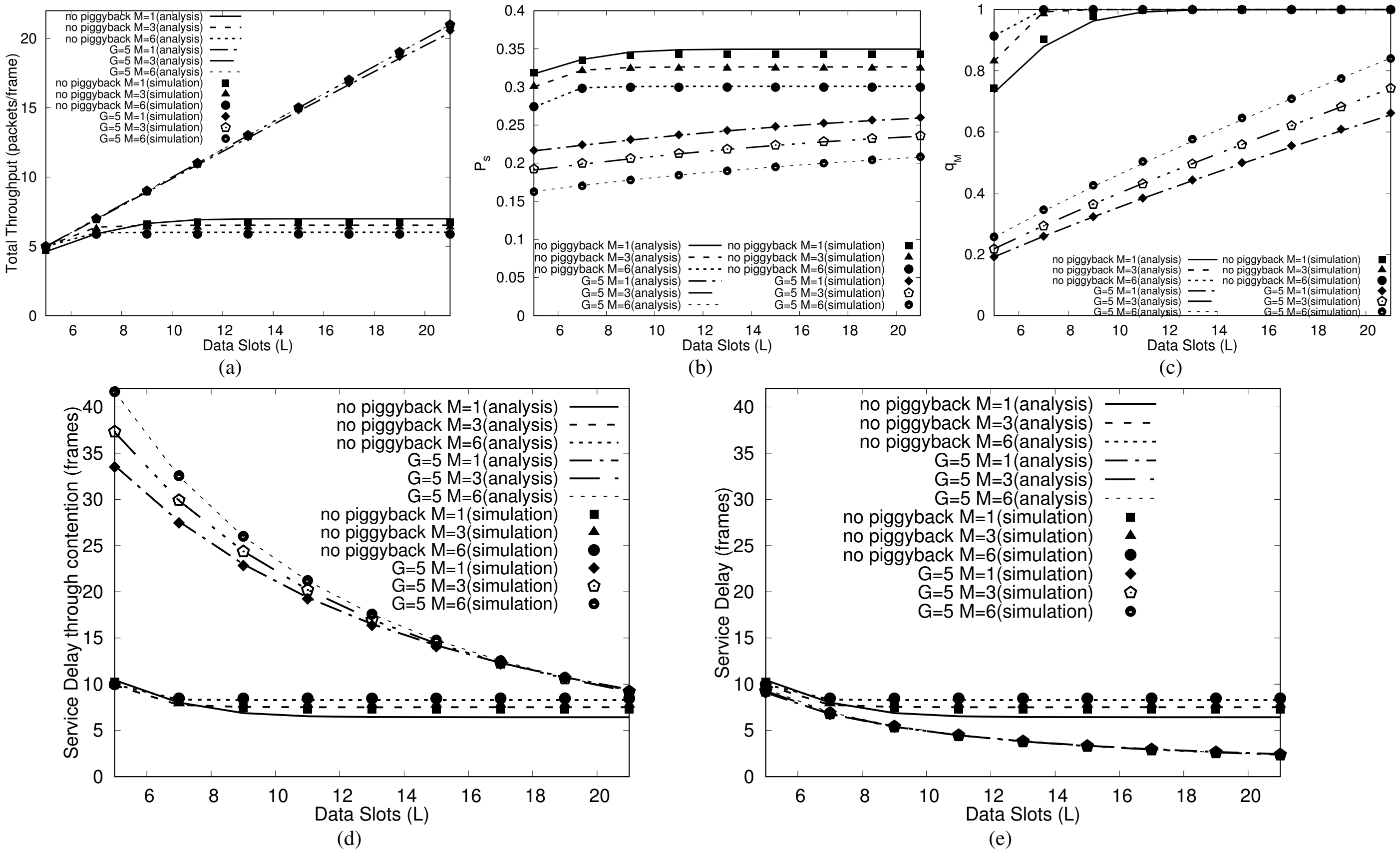
[23] J. Liu, S. Chan, X. Su, and H. L. Vu, "Performance analysis of contention based services with bulk transmission in ieee 802.16 ofdma networks," in *2014 IEEE 79th Vehicular Technology Conference (VTC Spring)*. IEEE, 2014, pp. 1–4.

[24] C. H. Wei, G. Bianchi, and R. G. Cheng, "Modeling and analysis of random access channels with bursty arrivals in ofdma wireless networks," *IEEE Transactions on Wireless Communications*, vol. 14, no. 4, pp. 1940–1953, April 2015.

[25] A. Vinel, Q. Ni, D. Staehle, and A. Turlikov, "Capacity analysis of reservation-based random access for broadband wireless access networks," *Selected Areas in Communications, IEEE Journal on*, vol. 27, no. 2, pp. 172–181, 2009.

[26] L. Kleinrock, *Queueing Systems*. Wiley Interscience, 1975, vol. I: Theory.

[27] R. B. Cooper, *Introduction to Queueing Theory*, 2nd ed. New York, NY: North-Holland, 1981.



(a)

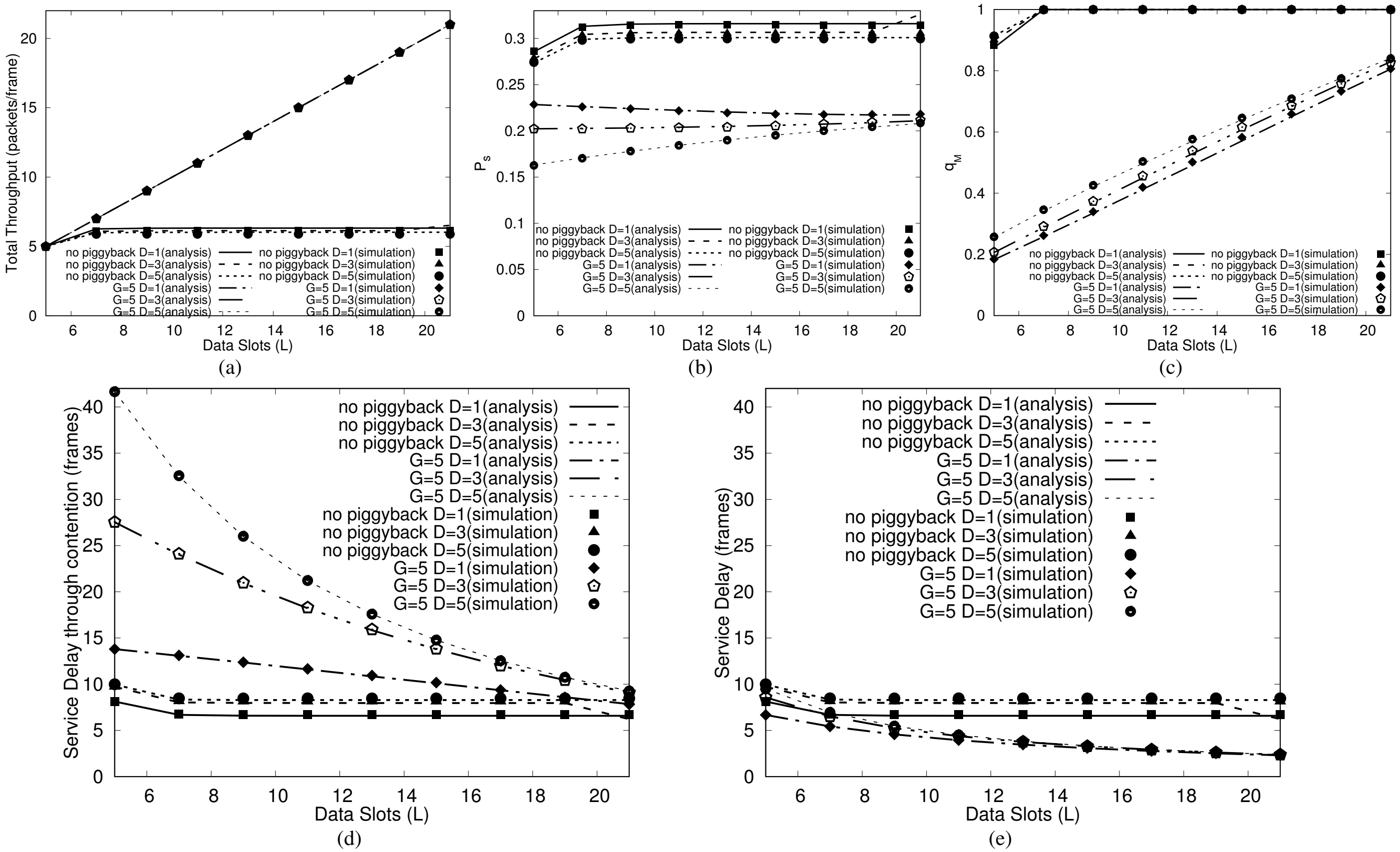
(b)

(c)

(d)

(e)

Fig. 11. Performance vs number of data slots (L) under saturated conditions ($\rho_{in} = 1$): (a) Total Throughput (Th) (b) P_s (c) q (d) Service Delay ($E\{S\}$) (e) Service Delay through contention ($E\{S_c\}$)



(a)

(b)

(c)

(d)

(e)

Fig. 12. Performance vs number of data slots (L) under saturated conditions ($\rho_{in} = 1$): (a) Total Throughput (Th) (b) P_s (c) q (d) Service Delay ($E\{S\}$) (e) Service Delay through contention ($E\{S_c\}$)