Performance Evaluation of Framed Slotted ALOHA with Reservation Packets for M2M Networks

Vicente Casares-Giner · Jorge Martinez-Bauset · Canek Portillo

Received: date / Accepted: date

Abstract In the near future, it is expected that a large number of machine-to-machine (M2M) communication devices will be part of ubiquitous information systems and help to develop multiple services. Random access protocols like ALOHA and CSMA have been considered for M2M networks for their simplicity of operation. This paper evaluates the performance of a Frame Slotted-ALOHA protocol that deploys reservation and data packets (FSA-RDP), in a scenario where a controller collect data packets transmitted by a finite number of M2M devices. In FSA-RDP, frames of variable duration are divided in two parts, the reservation and data subframes. During the reservation subframe, active devices send short reservation packets to the controller. The controller assigns reserved slots in the data subframe to those devices that succeeded with the reservation. At devices, the FIFO service discipline and two queue management schemes, tail drop and push-out, have been considered. When the queue size is of one packet, we develop a discrete-time Markov chain to evaluate the protocol performance, including the cumulative distribution function of the delay of data packets that are successfully transmitted. Analytical results are validated by extensive simulations. In addition, the simulation model is used to evaluate the system performance when larger queues are deployed. Numerical results show that the protocol efficiency of FSA-RDP is between one and two orders of magnitude larger than the efficiency of conventional Frame Slotted ALOHA. Also, we show that the difference between the packet delay for the tail drop and push-out schemes is only significant in scenarios with high load and high collision rate.

Keywords Machine-to-machine (M2M) communications · Frame Slotted ALOHA · discrete-time Markov chain · performance evaluation

1 Introduction

A fundamental part of the Internet of Things (IoT) is the concept of machine-to-machine (M2M) communications, that allows the autonomous exchange of data between a large population of devices [1]. The rapid increase in the number of M2M devices deployed brings serious design challenges. One of these challenges is the efficiency of the Medium Access Control (MAC) protocol. The operation simplicity of random access protocols, such as ALOHA or Carrier Sense Multiple Access (CSMA), makes them a good choice for M2M communications.

Since the seminal work of the ALOHA protocol [2], and its slotted version S-ALOHA [29], many different extensions have been proposed. A variant named Frame Slotted-ALOHA (FSA), studied in detail in [33][41], is worth mentioning. Due to its simplicity of implementation, FSA is being deployed in different practical systems, such as Radio Frequency Identification (RFID) [12], low power wide area networks (LPWAN) [18], or M2M communications [39][17], to name just a few examples.

A common operation feature in many of these systems is that FSA is used to transmit only data packets. However, when a collision occurs the entire data slot is lost. A more efficient approach is to deploy short reservation packets (RVP) for contention, and allocate reserved data slots for those devices that succeeded with
the reservation (their reservation packets did not collide). In this way, when RVP collide, only reservation mini-slots are lost, instead of the more valuable data slots. We call this protocol FSA with reservation and data packets (FSA-RDP).

FSA-RDP is different to other S-ALOHA reservation protocols like R-ALOHA [10]. R-ALOHA was proposed for scenarios where devices have to transmit long messages, that are divided into shorter packets. The contention is solved using data (packet) slots. Once a device successfully sends the first packet of the message, data slots are reserved in subsequent frames for the transmission of the rest of the packets that compose the message. Then, FSA-RDP could also be used to improve the performance of R-ALOHA.

The main contribution of this paper is the analytical performance evaluation of the FSA-RDP protocol. For the M2M devices, we consider the FIFO (first in, first out) service discipline with two queue management schemes, tail drop (TD) and push-out (PO) [11]. When TD is deployed, a new data packet is lost if it arrives to a M2M device with a full buffer. However, when PO is deployed, the arrival of a new data packet pushes-out the data packet stored at the head of the full buffer. For M2M devices with a buffer size of one packet we develop a discrete-time Markov chain (DTMC) to determine the packet delay distribution and several performance parameters. A preliminary analytical study of the FSA-RDP protocol was presented in [8]. An extension of the previous study for buffer sizes larger than one is also provided, where performance values are obtained by simulation.

The paper is organized as follows. After the introduction, Section 2 provides a comprehensive overview of past ALOHA contributions. We highlight the merits and shortcomings of previous proposals. A description of our contributions and the motivations of our work is presented in Section 3. Section 4 describes the basic operation of the system under study, and introduces the DTMC that defines the evolution of the number of active devices with time. In Section 5 the probability generating functions of the delay distribution for successfully transmitted packets are derived, both for the TD and PO queue management schemes. In Section 6 the analytical model is validated by extensive simulations. Also, a study of the efficiency of the FSA-RDP protocol is proposed. The efficiency is measured in terms of the packet loss probability, and compared to the one obtained by the conventional FSA. Results for the channel utilization, cumulative distribution function of the packet delay, average delay, 95-th percentile of the packet delay are also obtained when devices have buffer sizes of one, five and ten packets. The conclusions of the study are described in Section 7.

2 Related Work

Random Access Protocols (RAP) are commonly implemented in wired and wireless access networks. RAP are inherently unstable, so stabilization procedures must be seriously considered before being deployed in practical implementations. The stabilization procedure for RAP is implemented by combining two types of protocols or algorithms, channel access algorithms (CAA) and contention resolution algorithms (CRA) [38]. According to [16], they can also be characterized as global and local strategies, respectively. While CAA define rules by which new packets access the channel, CRA indicate how collisions are resolved. Most of the known CAA and CRA appear in isolation, although sometimes they join forces with each other [23].

Three main CAA categories have been considered in the literature. One is the free channel access, in which new packets are immediately transmitted in the first slot following their arrival. They are sometimes referred to as immediate first transmission (IFT). In other words, new and backlogged packets (those retrying) are treated identically [28].

A second category is the gated channel access, in which new arrivals delay their first transmission attempt until the contention resolution interval (CRI) during which they arrived has ended. In [22][31] they are referred to as obvious channel access algorithm. A pioneer example is the one proposed in [6].

A third category is the window channel access, in which time is divided into consecutive intervals of $\Delta_i$, seconds, $i = 1, 2, \ldots$. Packets that arrive during $\Delta_i$ produce the $i$-th CRI, i.e., they access during that CRI. The $\Delta_i$ time intervals can be adapted to traffic conditions. A window with constant duration, $\Delta_i = \Delta \forall i$, has also been adopted [27][26][3]. Both, gated channel access and window channel access estimate the number of active devices based on the access outcome in the previous access period. They are also collectively referred to as blocked channel access [23][38], or delayed first transmission (DFT) [28].

Two main CRA have been treated extensively in the literature. First, splitting algorithms are based on splitting a given conflict into smaller ones. Splitting algorithms offer different implementations, see [34], such as the stack algorithm [36][6][26], and the interval searching algorithm, in which the splitting is done based on packet arrival times [13][37].

Second, another approach is to treat all devices involved in a conflict identically. Within this second sub-
set of algorithms we can find the network centralized control by Bayesian broadcast probability proposed in [28]. A similar approach is used in [7], where a common transmission or permission probability, estimated similarly as in [28], is used by all devices involved in the same CRI. While in [28] the adopted CAA follows the IFT principle, the one proposed in [7], that is formally studied in [3], is a window channel access scheme and then follows the DFT principle.

A common feature of [28] and [7][3], is that the estimated number of packets in conflict is updated in an additive manner. On the contrary, the exponential increment/decrement action after the observation of consecutive collision and idle slots is implemented in Q-ary tree algorithms (stack algorithms), and in some of its enhanced versions, such as the Q+ary algorithm [15][21]. See also the multiplicative scheme in [42].

The maximum stable throughput reachable by a RAP depends on the stabilization procedure deployed. The system throughput is defined as the expected number of successful packets transmitted per time-slot. To evaluate the throughput, as well as the delay of a RAP, several aspects must be taken into account. For example, channel characterization, conflict feedback, full or limited feedback when sensing the channel, packet arrival process and buffer size per device.

Most of the proposed analytical models assume an ideal wireless channel with no errors, negligible propagation delay, and no capture effect (when a collision occurs, two or more devices are involved). Also, an instantaneous ternary (E/S/C; E:empty, S:success, C:collision) or binary (C/NC or S/NS, NC:No C, NS:No S) feedback is quite often assumed [38]. With respect to the arrival process, for tractability, the standard approach is to assume a Poisson (Bernoulli) arrival process for infinite (finite) number of devices. However, some studies deploy different arrival processes, such as D-BMAP. A detailed study for a family of Q-ary splitting algorithms is reported in [38], where packets arrive according to a discrete-time batch Markovian arrival process (D-BMAP) [5].

In the following examples it is assumed that the channel capacity is unity, and packets fit into a single slot. Then, the throughput expressed in packets per time-slot and the channel utilization numerically coincide. Also, it is assumed that the arrival process is Poisson with rate λ packets/slot. In [28], with E/S/C feedback and with the IFT policy, the stability of the RAP is guaranteed for λ \( \leq \frac{1}{e} \approx 0.3678 \). The splitting algorithm proposed in [6] considers a window channel access (CAA with DFT), also referred to as dynamic binary tree algorithm (DBT). It is stable for arrival rates \( \lambda < 0.4295 \), and with optimum constant window size \( \Delta^* = 2.677 \). In [26] a feedback with C/NC was proposed and evaluated. It is stable for \( \lambda < 0.4295 \) and with optimum window size \( \Delta^* = 2.33 \), i.e., the same throughput obtained in [6], but now with lower delays. Using the assumptions of [26], an analytical model for the proposal in [7] is developed in [3]. It is stable for \( \lambda < 0.4335 \) and with optimum window size \( \Delta^* = 2.49 \).

2.1 Frame Slotted ALOHA

Frame Slotted ALOHA (FSA) is an important family of RAP. FSA can be seen as a generalization of the standard S-ALOHA. In FSA a set of consecutive F > 1 slots conform one frame. At the beginning of a frame, each device that has a packet to transmit selects one of the F time-slots with equal probability. If the device does not receive an acknowledgment from the central controller, it will try again in the next frame. FSA and its variants have been extensively evaluated in the literature.

One example is the study in [41], where an exact performance evaluation analysis of FSA with capture is presented. The study assumes a finite number of unbuffered devices (i.e., can hold at most one packet at a time), and a Bernoulli traffic model. Another example is the work in [32], where a dynamic FSA (DFSA) is proposed. Here, the frame length F is estimated according to the expected value of the number of backlogged data packets at the beginning of a frame. Under the assumptions of a Poisson arrival process, single data packet per message, noise-less channel and no capture effect, the proposed DFSA achieves a maximum throughput of 0.426. Clearly, it is slightly higher than the one achieved by an stabilized S-ALOHA (1/e \( \approx 0.3678 \)).

We highlight the fact that the maximum achievable throughput in FSA with infinite number of sources, Poisson arrival process, and without dynamic frame length adaptation, equals the one achieved by the ordinary S-ALOHA. It is also worth mentioning that FSA has been adopted by RFID systems [12].

2.2 Reservation Schemes

S-ALOHA and FSA protocols can be used to transfer messages composed by multiple packets. One pioneering example is R-ALOHA proposed in [10] to improve the throughput of satellite communications. Here, the contention is solved using data (packet) slots. Once a device successfully sends the first packet of the message, data slots are reserved in subsequent frames for the transmission of the rest of the packets that compose the message.
An improved protocol was proposed in [30], where frames are divided into contention mini-slots and data time-slots. Mini-slots are deployed for the out-of-slot signaling protocol (reservation), in opposition to in-slot signaling protocols where contention occurs in data time-slots. Once a device succeeds with a reservation, it can transfer a packet in a reserved data time-slot, or a full message in reserved data time-slots in subsequent frames. Note that even with in-slot signaling, the effective throughput of the channel can increase far beyond the maximum achievable by ordinary S-ALOHA ($1/e \approx 0.3678$).

Since then, extensive studies of reservation FSA (RFSA) have been conducted. Some representative examples are [33][19][35]. In [33], it is assumed that devices have buffers that hold a single packet, packets arrive according to a Bernoulli model, and out-of-slot signaling is deployed. Successful request are placed in a common FIFO queue of finite capacity. The study provides an exact numerical analysis and computable formulas for delay and throughput.

In contrast, in [19] and [35] it is assumed that devices are equipped with buffers that hold a single multi-packet message, messages arrive according to a Bernoulli model, and in-slot signaling is deployed. At frame initiation, devices with a message to transmit randomly choose one of the non-reserved data slots. If the device succeeds with the transmission of the first packet, slots will be reserved to transmit the rest of the packets of the message in the following frames. Otherwise, it contends in the next frame. Clearly, these protocols operate with the IFT principle.

An interesting reservation protocol is the Packet Reservation Multiple Access protocol (PRMA). PRMA was one of the pioneering applications of multi-access protocol to support real-time services. It was proposed in the late 1980s and developed in a series of papers [14][24]. PRMA systems incorporate speech activity detectors, and operate with the idea that the front-end of a train of voice packets that conforms a talkspurt is in charge of channel reservation. The first packet of a voice talkspurt is lost if it collides with other voice packets. Then, the next voice packet of the same talkspurt will take care of the reservation for the rest of the talkspurt. PRMA follows a loss model, as a clipping of the front-end might occur due to collisions.

RFSA has also been proposed and studied in [40][39] for M2M applications. In [40] a finite number of devices each one with a multi-packet message ready to transmit try to gain access by means of the first data packet and according to the FSA protocol. Devices that successfully access, get a reserved slot in order to transmit the rest of the message in subsequent frames. Note that the protocol is very similar to the PRMA, but for messages that are delay insensitive.

3 Motivation and Contributions of the Study

We propose an analytical model to evaluate the performance of a FSA scheme in which mini-slots are used for reservation and data slots to transmit data packets (out-of-slot signaling). Frames are composed of a reservation subframe (RSF) and a data subframe (DSF). Typically, the RSF is divided into multiple mini-slots, that devices use for contention access following the FSA scheme. The DSF is composed of a variable number of data slots, the duration of each of them being multiple mini-slots. Once a device succeeds with a reservation, it transmits a single data packet in one data slot of the same frame. Then, the maximum number of data slots in a frame equals the number of reservation mini-slots. Note that occasionally it may happen that the DSF has zero data slots. Clearly, devices operate following the IFT principle.

Packets arrive to devices according to any renewal arrival process, as the model characterizes the number of packets that arrive to a device per mini-slot by independent and identically distributed random variables. Then, at each mini-slot, packets arrive following a general distribution irrespective of where the packets were originated, i.e., internally (as a consequence of its own sensing activity) or externally (arriving from another device). For simplicity, we assume that packets arrive according to a Poisson process with a constant rate. Therefore, the number of packet arrivals per mini-slot follows a discrete Poisson distribution.

As in [33], for model tractability, we assume that devices are equipped with buffers that hold a single data packet. However, we analyze systems with buffer sizes larger than one by simulation. Also for tractability, we consider that the channel is noise-less, with no capture effects, and instantaneous ternary feedback.

To the best of authors knowledge, the proposed model is the first analytical model proposed in the literature to evaluate the performance of the Frame Slotted ALOHA with Reservation and Data Packets (FSA-RDP) at the mini-slot granularity. One of the main differences between the model proposed in [33] and the one proposed in our study is that the duration of the frame is variable, instead of being of fixed duration as in [33]. This is an important difference as it leads to a higher channel utilization. In addition, we consider two queue management schemes, tail drop (TD) and push-out (PO). We believe that PO schemes have an important role in scenarios where sensors report measurements of the
environment. Here, the last measurement might be of higher value than past ones.

Finally, our analytical approach is completely different to the one in [33]. We exploit the more structured matrix geometric analysis, while the model in [33] is based on conventional probabilistic arguments. In addition, the proposed model allows to determine the packet delay distribution at the granularity of the mini-slot, while the model in [33] allows only to determine average delays at the granularity of the frame duration.

An ALOHA protocol that deploys reservation mini-slots and data slots has been studied in [20] by simulation. However, FSA-RDP is different to the protocol studied there, where no packets are allowed to enter the system until the controller fully resolves the collisions that occurred during the last access, deploying for this a distributed queue algorithm. That is, a DFT approach, in contrast to the IFT deployed in our work.

4 System model

4.1 Protocol Operation

We consider a wireless network composed by a controller (gateway) and a finite number of devices forming a star topology. Each device operates independently and transmits data packets (DAP) of constant size as a consequence of its activity. In the RSF, active devices (with packets in their queues) contend for channel access by transmitting reservation packets (RVP) to the controller. At the end of the RSF, the controller broadcasts a slot allocation packet (SAP) that contains the data slots allocated to those devices that succeeded with the reservation. Following the SAP, successful devices initiate the transmission of one DAP per reserved data slot, i.e., in a contention free manner. Finally, the controller broadcasts an control packet to acknowledge the successful reception of the data packets.

Figure 1 describes an example of the operation of FSA-RDP, where 5 devices transmit packets to the controller. Observe that we denote by: i) Di the devices; ii) Ri the RVP transmitted by Di; iii) Data i the data packet transmitted by Di; iv) S successful RVP; v) C collided RVP; vi) E empty reservation mini-slot. At frame i − 1 four devices contend for access sending their respective RVP during the RSF. D2 and D5 succeed, while D1 and D4 collide. At the end of the RSF, the controller broadcasts the SAP to allocate data slots to the successful devices. Then, D2 and D5 send a single DAP during the DSF. At frame i, D1 and D4 collide again. Then, the controller broadcasts the SAP informing that the DSF is omitted in this frame. At frame i + 1, D1 and D3 collide while D4 succeeds. Then, the DSF ends after D4 transmits a single DAP. Observe that D3 was inactive at the beginning of frame i − 1 and frame i, and received a packet during frame i.

4.2 Network Model

We consider a network composed by a finite number of M devices that transmit packets to a central controller (gateway). For model tractability, packets arrive to a device according to a Poisson process with rate λ packets/mini-slot, and are equipped with a buffer that can hold a single packet. Let Q denote the buffer size in packets, then Q = 1.

Let W be the duration of a data slot in mini-slots. Also, let V be the number of reservation mini-slots in the RSF. It can be shown that the reservation success probability is maximized when the number of contenders in each RSF is 

\[ n_{\text{opt}} = \left( \frac{\ln(V/(V - 1))}{V} \right)^{-1} \]

From the basic inequalities, \( 1 - 1/x < \ln(x) < x - 1 \) for \( x > 0 \), it follows immediately that \( V - 1 < n_{\text{opt}} < V \). This suggest than the number of reservations mini-slots V per frame should be adapted to the number of estimated contenders. However, for tractability, we consider a fixed value for V.

We introduce an access permission probability r. Then, at each RSF, active devices send a RVP with probability r. Note that r can be adjusted by the controller according to the observed outcomes of the previous RSF and the estimated arrival traffic during the whole frame. The adaptive algorithm is left for further study. Then, we consider that the value of r is fixed.

Let i be the number of active devices at the beginning of a RSF. The probability that j of them, \( j \leq i \leq M \), transmit a RVP follows a binomial distribution, and denote it as \( B_j \sim (r) \). Note that a device selects any of the V mini-slots to transmit the RVP with equal probability (1/V).

Let \( S_k^j(V) \) denote the probability that k among j contending devices succeed with the reservation. It can
be obtained recursively as,
\[
S^i_k(V) = \sum_{n=0, \neq 1}^j B^j_n(1/V)S^i_{k-n}(V-1), \quad k = 0, \\
S^i_k(V) = B^j_1(1/V)S^i_{k-1}(V-1) + \sum_{n=0, \neq 1}^j B^j_n(1/V)S^i_{k-n}(V-1), \quad k = 1, \ldots, V.
\]

Recurrence (1) is based on the following sequential reasoning. First, we observe if the contention outcome in the first mini-slot of the RSF is: empty (no RVP), success (a single RVP), or collision (multiple RVP).

Based on this outcome, we formulate the contention outcome in the remaining mini-slots of the RSF.

Let \( D_k^i(r, V) \) denote the probability that \( k \) among \( i \) active devices succeed in a frame. It is given by,
\[
D_k^i(r, V) = \sum_{j=k}^i B_j^i(r)S^i_j(V), \quad 0 \leq k \leq j \leq i.
\]

The evolution of the number of active devices observed at the beginning of each RSF can be modeled as a DTMC. Let \( \{P_{i,j}\} \) denote the transition probabilities of the DTMC, i.e., the probability of \( j \) active devices in the next frame \( n+1 \), conditioned on \( i \) active ones in the current frame \( n \). They are given by,
\[
P_{ij} = \sum_{k=\max(0, i-j)}^{\min(i,V)} D_k^i(r, V)A_{j-i+k}^M(a_k),
\]

where \( A_n^s(a_k) \) is the probability that \( n \) devices become active among \( s \) inactive in a frame with \( k \) successful reservations. For Poisson arrivals \( A_n^s(a_k) \) follows a binomial distribution, where \( a_k = 1 - e^{-t_k} = 1 - \delta^k \), \( t_k = V + kW \), for \( 0 \leq k \leq V \) (for convenience, we define \( \delta = e^{-\lambda} \)), and \( k \) is number of successful reservations in the frame. Then, \( a_k \) is the probability that at least one data packet arrives to a device during a frame of duration \( t_k \) mini-slots.

Note that the definition of \( t_k \) allows the model to determine the delay of data packet at the granularity of a mini-slot. Also, notice that since we assume \( Q = 1 \), no more than one data packet per frame and per device can be admitted. Finally, we point out that both, the transmission of a SAP at the end of a RSF, and the transmission of an ACK at the end of a DSF by the controller, require a single mini-slot. For simplicity, this constant delay term has been omitted in the analysis.

Then, with TD scheme, the first data packet that arrives along frame \( n \) (the new data packet) is transferred to the buffer for transmission if it is found empty. If the buffer is occupied with a previous data packet that succeed in the RSF \( n \), the new data packet also joins the buffer. Otherwise, the new data packet is dropped. However, with PO, the last data packet that arrives along a frame \( n \) (the new data packet), is transferred to the buffer for transmission regardless of the state of the buffer. In particular, if the buffer is found occupied with a previous data packet, and this data packet did not succeed in the RSF \( n \), it is pushed-out by the new one.

Let \( \pi \) denote the stationary probability of finding \( i \) active devices at the beginning of an arbitrary RSF. The stationary probabilities \( \pi = [\pi_0, \pi_1, \ldots, \pi_M] \) are obtained by solving the system of linear equations \( \pi = \pi P \) with the normalization condition of \( \pi e = 1 \), where \( P = [P_{ij}] \), and \( e \) is a column vector of 1s. Then, the throughput \( \gamma \) (carried data packet rate), in packets per mini-slot, can be evaluated as,
\[
\gamma = \sum_{k=0}^V k f_k / \sum_{k=0}^V t_k f_k, \quad f_k = \sum_{n=k}^M \pi_n D_n^k(r, V),
\]
where \( f_k \) is the fraction of frames which duration is equal to \( t_k = V + kW \).

5 Data Packet Delay Distribution

Let us consider a frame randomly chosen, frame \( n \), and suppose that its duration is \( t_m = V + mW \). We assume that our tagged device receives at least one data packet during that frame; this occurs with probability \( a_m \). Note that only one packet can be hold at the device buffer (\( Q = 1 \)). We observe the number of active devices at the beginning of that frame. A device is active when has a packet in its buffer. Clearly, since the frame has a duration \( t_m \), at least \( k \) devices, \( m \leq k \leq M \), were active at the beginning of that frame \( (M - k) \) were inactive). Our tagged device could be one of the active devices or not. In the following Subsections we distinguish three different cases.

5.1 Case A

At the beginning of frame \( n \), the tagged device is not active, i.e., it is not among the \( k \) active ones. This occurs with probability \( (M - k)/M \). Assuming that \( m \) out of \( k \) terminals gain access in the RSF \( n \), the probability that our tagged terminal will compete in the next frame \( n + 1 \) with other \( j \) terminals, \( 0 \leq j \leq M' = M - 1 \), is given by,
\[
M - k \frac{M - k - j}{M} D_m^j(r, V)A_j^{M'-k+m}(a_m).
\]
Weighting the above expression by the steady state probabilities $\pi_k$, $m \leq k \leq M$, we have,

$$P_A(M, m, j) = \sum_{k=m}^{M} \pi_k \frac{M - k}{M} D_m^k(r, V) A_{j-k+m}^{M-k+m} (a_m).$$  \hspace{1cm} (5)

Since the probability that our tagged terminal receives at least one data packet during the frame $n$ ($a_m$) depends on the frame duration $t_m$, (5) must be weighted by $a_m$. Furthermore, in order to include the arrival instant of the packet within frame $n$ we have to distinguish between the TD and the PO schemes. In particular, respectively we have,

$$F_m(z) = \sum_{l=0}^{t_m-1} \delta^{l-1}(1-\delta)z^l, \hspace{1cm} (6)$$

$$L_m(z) = \sum_{l=0}^{t_m-1} (1-\delta)\delta^l z^l. \hspace{1cm} (7)$$

Both expressions (6) and (7) are generating functions that account for the admitted data packet arriving at any mini-slot of the frame $n$. Obviously, $a_m = F_m(z=1) = L_m(z=1)$. They are explained as follows. In (6), we keep in mind that the duration of frame $n$ is $t_m$ mini-slots. Then, for $l=0$, a data packet arrives to the tagged device during the last mini-slot of frame $n$, and no data packets arrive during the previous mini-slots. This occurs with probability $\delta^{t_m-1}(1-\delta)$. For $l = 1$, a data packet arrives to the tagged device in the penultimate mini-slot of frame $n$, and no data packets arrive during the previous mini-slots. This occurs with probability $\delta^{t_m-2}(1-\delta)$. Note that the potential arrival of one data packet during the last mini-slot has no effect, as this data packet is rejected because the TD scheme is being considered. For $l = 2, 3, \ldots, t_m - 1$ the same reasoning applies.

For expression (7) we argue as follows. For $l = 0$, with probability $(1-\delta)$ a data packet arrives to the tagged device during the last mini-slot, and pushes-out previous arrivals, if any. For $l = 1$, with probability $(1-\delta)$ a data packet arrives to the tagged device during the penultimate mini-slot and pushes-out previous arrivals, if any. It will contend in the RSF $n+1$ if no arrivals occur during the last mini-slot. This occurs with probability $(1-\delta)\delta$. This line of reasoning can be extended to the rest of mini-slots.

Also, the normalization with respect to the frequency of occurrences of intervals of duration $t_m$ weighted by $a_m$ is required. Then, we can write,

$$P_{A,X}(M, j, z) = (1/G_a) \sum_{m=0}^{V} X_m(z) P_A(M, m, j), \hspace{1cm} (8)$$

where $G_a = \sum_{i=0}^{V} a_i f_i$ and $X = F, L$; i.e. $X_m(z) = F_m(z), L_m(z)$, for TD or PO schemes, respectively.

5.2 Case B

At the beginning of the frame $n$ the tagged device is active, i.e., it is one of the $k$ active devices. This situation occurs with probability $k/M$. Assuming that $m$ out of $k$ devices succeed with the reservation, the tagged device succeeds with probability $m/k$. Then, the packet preceding our tagged packet is successfully transmitted in the actual frame $n$. The tagged packet is copied to the buffer as soon as the previous packet in the buffer is transmitted. In that case, the probability that our tagged device will compete in the next frame $n+1$ with other $j$ active devices, $0 \leq j \leq M’ = M - 1$, is given by,

$$k \frac{m}{k} D_m^k(r, V) A_{j-k+m}^{M-k+m} (a_m).$$

As in (5), weighting the above expression by the steady state probabilities $\pi_k$, $m \leq k \leq M$ we have

$$P_B(M, m, j) = \sum_{k=m}^{M} \pi_k \frac{m}{M} D_m^k(r, V) A_{j-k+m}^{M-k+m} (a_m). \hspace{1cm} (9)$$

Finally, in the same way as in (8), weighting the expression (9) we get,

$$P_{B,X}(M, j, z) = (1/G_a) \sum_{m=0}^{V} X_m(z) P_B(M, m, j). \hspace{1cm} (10)$$

5.3 Case C

This case complements Case B, where we assumed that the tagged device was active. In addition, we assume here that the tagged device does not succeed with the reservation in frame $n$. This occurs with probability $(k-m)/k$. Then the tagged data packet that arrives along frame $n$ will be lost with the TD scheme ($Q = 1$), but admitted with the PO scheme (as it pushes-out the packet in the buffer). In this later case, the probability that the tagged packet contends with other $j$ active devices in frame $n+1$, $0 \leq j \leq M’ = M - 1$, is given by,

$$\frac{k}{M} k \frac{m}{k} D_m^k(r, V) A_{j-k+m+1}^{M-k+m+1} (a_m).$$

Then, as in (5) and in (9), weighting the above expression by the steady state probabilities $\pi_k$, $m \leq k \leq M$.
\[ P_C(M, m, j) = \sum_{k=m}^{M} \frac{\pi_j}{M} D^k_m(r, V) A^{M-k+m+1}(a_m). \]  
\( (11) \)

In the same way as in (8) and in (10), weighting the expression (11) we get,

\[ P_{C,PO}(M, j, z) = \frac{1}{G_a} \sum_{m=0}^{V} L_m(z) P_C(M, m, j). \]  
\( (12) \)

5.4 Phase-Type Distributions

The delay distribution of successfully transmitted data packet under the TD and PO buffer management scheme follow Phase-type (PH) distributions. They are represented as \((\alpha, T, T^0)\), where \(\alpha\) defines the probabilities that the absorbing process is initiated at each of its transient states, \(T\) defines the transition probabilities between transient states, and \(T^0 = e - Te\) is a column vector that defines the transition probabilities from transient states to the absorbing state [25] [4].

For the TD and PO schemes we denote them as \((\alpha_{TD}, T_{TD}, T^0_{TD})\), and \((\alpha_{PO}, T_{PO}, T^0_{PO})\).

The initial probability vectors for the PH distributions follow from (8), (10) and (12). For TD we have, \(\alpha_{TD}(z) = [\alpha_{TD,0}(z), \cdots, \alpha_{TD,M'}(z)]\), and for PO scheme, \(\alpha_{PO}(z) = [\alpha_{PO,0}(z), \cdots, \alpha_{PO,M'}(z)]\), where

\[ \alpha_{TD,j}(z) = P_{A,TD}(M, j, z) + P_{B,TD}(M, j, z), \]
\[ \alpha_{PO,j}(z) = P_{A,PO}(M, j, z) + P_{B,PO}(M, j, z) + P_{C,PO}(M, j, z), \]

where \(\alpha_{TD}(1)e < 1\), i.e., \(\alpha_{TD}(1)\) is a substochastic vector, and \(\alpha_{PO}(1)\) is a stochastic vector, \(\alpha_{PO}(1)e = 1\).

5.5 Matrix \(T_{TD}|T^0_{TD}\) for TD scheme

Let \(T_{TD,k,i,j}\) be the probability that the tagged device contends with other \(i\) active devices in a frame, \(k\) devices succeed with the reservation, and the tagged device does not succeed, and contends in the next frame with \(j\) other active devices. Then, for \(0 \leq i, j \leq M' = M - 1\) and \(0 \leq k \leq i + 1\),

\[ T_{TD,k,i,j} = (1 - \frac{C_i}{C_k}) D^{i+1}_k(r, V) A^{M-i-1+k}(a_k) \]
\[ = (1 - \frac{k}{i+1}) D^{i+1}_k(r, V) A^{M-i-1+k}(a_k) \]
\( (13) \)

where \(C_i = (i)\). In matrix notation and in terms of the probability generating function (pgf) we have,

\[ T_{TD}(z) = \sum_{k=0}^{V} T_{TD,k,i} z^{(V+kW)}. \]  
\( (14) \)

To get a better understanding of (14) we remember that, when \(k\) devices succeed with the reservation, the RSF is followed by a DSF of duration \(kW\) mini-slots.

Let \(T_{TD,k,i}^0\) be the probability that the tagged device contends with other \(i\) active devices in a frame, \(k\) devices succeed with the reservation, and the tagged device is one of them. It is given by,

\[ T_{TD,k,i}^0 = \frac{C_i}{C_k+1} D^{i+1}_k(r, V) = \frac{k}{i+1} D^{i+1}_k(r, V). \]  
\( (15) \)

Let \(T_{PO,k,i}^0\) denote the column vector with elements given by (15). Again, in matrix notation and in terms of the generating functions, we have the following column vectors,

\[ T_{PO}(z) = \sum_{k=0}^{V} T_{PO,k,i}^0 z^{(V+kW)}. \]  
\( (16) \)

In the above expression, it is assumed that the controller will send the ACK at the end of the DSF to acknowledge the successful reception of data packets to the \(k\) corresponding devices. Note that once the tagged packet is successfully transmitted in a given frame, the arrival of new data packets has no impact on the tagged packet. That is, the term \(A^k_n(a_k)\) is no longer needed in (15), but was needed in (13).

5.6 Matrix \(T_{PO}|T^0_{PO}\) for PO scheme

Let \(T_{PO,k,i,j}\) be defined as (13) but with the additional condition that no packets will arrive to the tagged device during a frame of duration \(t_k\). Then,

\[ T_{PO,k,i,j} = (1 - a_k) T_{TD,k,i,j}. \]  
\( (17) \)

Let \(T_{PO,k}\) denote a matrix with elements given by (17). In matrix notation and in terms of the pgf we have,

\[ \sum_{k=0}^{V} T_{PO,k} z^{(V+kW)} = \sum_{k=0}^{V} (1 - a_k) T_{TD,k} z^{(V+kW)}. \]  
\( (18) \)

With regard to \(T_{PO,k,i}^0\), the PO counterpart of (15), we have two contributions. The first one contains the probability that the tagged data packet is pushed-out in
a frame of duration \( t_k \), before it succeeds with a reservation. This happens with probability \( a_k \sum_{j=0}^{M} T_{TD,k;j} = a_k T_{TD,k,i} \). The second contribution is coincident with the TD scheme, so with \( T_{TD,k}^0 \) given in (15), and reflects the probability that the reservation packet of the tagged device does not collide, and the corresponding data packet is successfully transmitted in a frame of duration \( t_k \). Then,

\[
T_{PO,k,i}^0 = T_{TD,k,i}^0 + a_k T_{TD,k,i},
\]

and the parallel expression to (16) is

\[
T_{PO}^0(z) = T_{TD}^0(z) + \sum_{k=0}^{V} a_k T_{TD,k}z^{(V+KW)},
\]

### 5.7 Data packet delay distributions

Based on previous section, here we derive the pdf of the delay distribution of transmitted packets for TD and PO schemes. Using the results given in (14) and (16), (18) and (20), then generating function of the sojourn time for TD is given by

\[
R_{TD}(z) = \alpha_{TD,M} + \alpha_{TD}[I + T_{TD}(z) + T_{TD}^2(z) + \ldots]T_{TD}^0(z) = R_{TD,M} + \alpha_{TD}[I - T_{TD}(z)]^{-1}T_{TD}^0(z)
\]

(21)

\( R_{TD,M} \) is the probability that a data packet is rejected by the tagged device, and \( R_{TD,a}(z) \) takes into account the delay distribution of the data packets accepted by the tagged device. Obviously \( R_{TD}(1) = 1 \).

For PO we have, with \( \alpha_{PO,M} = 0 \),

\[
R_{PO}(z) = \alpha_{PO}(z)[I + T_{PO}(z) + T_{PO}^2(z) + \ldots]T_{PO}^0(z) = \alpha_{PO}(z)[I - T_{PO}(z)]^{-1}T_{PO}^0(z)
\]

(22)

\[
= \alpha_{PO}(z)[I - T_{PO}(z)]^{-1}T_{TD}^0(z)
\]

\[= \sum_{k=0}^{V} a_k T_{TD,k}z^{(V+KW)}
\]

\[= R_{PO}(z) + R_{PO,a}(z),
\]

where \( R_{PO,M}(z) \) and \( R_{PO,a}(z) \) can be identified from (20). \( R_{PO,a}(z) = \alpha_{PO}(z)[I - T_{PO}(z)]^{-1}T_{TD}^0(z) \) takes into account the delay distribution of the data packets that are transmitted before being pushed-out. \( R_{PO,r}(z) \) takes into account the waiting time of those data packets that are pushed-out.

We are only interested in finding the delay for packets that are finally transmitted. That is,

\[
R_{TD,a}(z) = \alpha_{TD}(z)[I - T_{TD}(z)]^{-1}T_{TD}^0(z)
\]

(23)

\[
\text{Table 1: Model Symbols and Parameters}
\]

<table>
<thead>
<tr>
<th>Input Parameters</th>
<th>Symbol</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of devices</td>
<td>( M )</td>
<td>8</td>
</tr>
<tr>
<td>Numb. reserv. mini-slots</td>
<td>( V )</td>
<td>1, 2, \ldots</td>
</tr>
<tr>
<td>Permission probability</td>
<td>( \gamma )</td>
<td>[0, 1]</td>
</tr>
<tr>
<td>Device buffer size</td>
<td>( Q )</td>
<td>{1, 5, 10}</td>
</tr>
<tr>
<td>Data packet duration</td>
<td>( \lambda )</td>
<td>packet/mslot</td>
</tr>
<tr>
<td>Frame duration (mini-slots)</td>
<td>( t_k )</td>
<td>( V + KW )</td>
</tr>
<tr>
<td>Device arrival rate</td>
<td>( \lambda )</td>
<td>packet/mslot</td>
</tr>
<tr>
<td>Device load</td>
<td>( \rho_i )</td>
<td>( \lambda V )</td>
</tr>
<tr>
<td>System load</td>
<td>( \rho_t )</td>
<td>( M )</td>
</tr>
</tbody>
</table>

\[
R_{PO,a}(z) = \alpha_{PO}(z)[I - T_{PO}(z)]^{-1}T_{PO}^0(z)
\]

(24)

with \( R_{TD,a}(1) = R_{PO,a}(1) = 1 - \alpha_{TD,M} \).

### 6 Numerical results

Different performance parameters are studied, such as the protocol efficiency measured in terms of the data packet loss probability. Also, we determine the channel utilization, the packet throughput, the cumulative distribution function (CDF) of the packet delay for packets that are successfully transmitted, and from it the 95-th percentile of the packet delay.

We first validate the analytical model for \( Q = 1 \) by comparing its performance results with those obtained by simulation. The simulation results are obtained by implementing the contention schemes, scheduling schemes, queue management schemes, and data transmission procedure in a custom-built C based discrete-event simulation program. The developed simulation model mimics the physical behavior of the MAC protocol considered (FSA-RDP). That is, in each frame a device receives packets according to a given discrete distribution, contends for channel access with other devices using reservation packets when it is active (has packets in the buffer), and, if it succeeds, then transmits a data packet in the data slot allocated by the controller.

The simulation results are completely independent to those obtained by the analytical model. That is, the computation of performance metrics by the simulation
model is not dependent on the derived mathematical expressions at all, nor are the state transition probabilities used in these computations. The performance results reported are the average values of measurements made over 20·10^6 data packet arrivals. In addition, 95% confidence intervals have been obtained. However, as they are very small, and therefore not significant, they are only shown in Fig. 2 and 3 for illustration purposes, and omitted in the rest of figures.

We define a reference scenario where \( M = 8, W = 10, Q = 1 \), and permission probability \( r = 1 \), unless otherwise stated. The load of device \( i \) is \( \rho_i = \lambda_i W \), where \( \lambda_i \) is the device packet arrival rate in packets per mini-slot. Then, \( \rho_i \) is the average number of packets per data slot that arrive to device \( i \). The total traffic load is defined as \( \rho_T = \sum_{i=1}^{V} \rho_i \). Also, \( \lambda_T = \rho_T \). Three reference traffic loads are considered: i) low, \( \rho_T = 0.3 \); ii) medium, \( \rho_T = 0.5 \); and iii) high, \( \rho_T = \{0.8, 0.9\} \).

We compare the packet loss probability and the average packet delay to validate the analytical and simulation models for systems with \( Q = 1 \). According to (4), the data packet loss probability \( P_L \) is given by,

\[
P_L = 1 - \frac{\gamma}{M\lambda} = 1 - \frac{\gamma}{\lambda_T}.
\]  (25)

Note that for both queue management schemes, TD and PO, the fraction of data packets that are lost must be the same. When a packet arrives to a full buffer, one packet is lost regardless of the queue management scheme. With TD, the packet that arrives is lost, while with PO, the packet at the head of the queue is lost. This intuition is confirmed by the results of the analytical and simulation models.

Figure 2 shows the variation of the packet loss probability with the offered load \( \rho_T \), obtained by the analytical model (ana) and by simulation (sim), for \( V = \{2, 4\} \).

Results obtained by the analytical model are shown in green, while results obtained by the simulation model are shown in black. Devices deploy the FIFO queue discipline with either tail drop (TD) or push-out (PO) queue management schemes. As observed, the analytical and simulation results match exactly.

Note that Fig. 2 displays a somehow counter intuitive system behavior. That is, for loads below \( \rho_T = 0.7 \) \( P_L \) for a system with \( V = 2 \) is lower than for as system with \( V = 4 \). This is due to the fact that average frame duration for \( V = 4 \) is longer than for \( V = 2 \). In longer frames, the probability that multiple arrivals occur during the frame is larger. As devices can only hold a single packet in their buffers, losses will more likely occur in longer frames. However, as load increases, allowing \( V = 2 \) reservation mini-slots instead of \( V = 4 \) becomes a bottleneck. As more collisions occur, packets are held longer in the buffer, and more losses occur induced by new arrivals. Recall that in FSA-RDP losses may only occur at new packet arrivals, as collisions in the channel occur with RVP.

Figure 3 shows the variation of the average packet delay with the offered load \( \rho_T \), obtained by the analytical model (ana) and by simulation (sim), for \( V = \{2, 4\} \). The packet delay is the time elapsed since packet arrival to departure (when the last bit is transmitted). Results obtained by the analytical model are shown in green, while the results obtained by simulation are shown in black (PO) and blue (TD) colors. As observed, the analytical and simulation results for the PO and TD queue management schemes match exactly.

Note in Fig. 3 that, as expected, the packet delay for PO is lower than for TD. When PO is deployed and a new packet arrives to a full buffer, the packet at the head of the queue is pushed-out. Clearly, the
Following with the case $Q = 1$, we compare three protocols: i) FSA; ii) FSA-RDP; and iii) ideal FSA-RDP (FSA-RDP-I). In FSA the contention occurs in data slots (in-slot signaling), while in FSA-RDP occurs in reservation mini-slots (out-of-slot signaling).

In FSA-RDP-I there is no contention. It is assumed that the controller knows the state of the devices at the beginning of a frame, and it sends the SAP without the need for contention. In order to perform a fair comparison, the frame structure for FSA-RDP-I is the same as the one defined for FSA-RDP in Fig. 1. That is, it is also composed by the RSF and DSF, and the maximum number of packets that can be transferred during the DSF is $V$. Therefore, the analytical model for FSA-RDP-I is the same as the one developed for FSA-RDP, except that the permission probability is set to $r = 1$, and $D_k^i(1,V)$ is now defined following (2) as,

$$ D_k^i(1,V) = \begin{cases} 1 & i \leq V, k = i, \\ 1 & V < i, k = V, \\ 0 & \text{otherwise} \end{cases} $$

Table 2 shows the minimum $P_L$ for different values of $V$ and for the three loads considered. Note that the minimum $P_L$ is achieved at the corresponding optimal permission probability ($r_{opt}$), that is obtained by exhaustive search. As expected, the performance of FSA-RDP is upper bounded by the performance of FSA-RDP-I as revealed by the results.

Observe that for a given $V$, $P_L$ is larger for FSA than for FSA-RDP. In FSA data packets losses occur mainly due to access collisions, and less frequently due to arrivals to a full buffer. However, in FSA-RDP access collisions occur only for RVP, and data packet losses are only due to arrivals to a full buffer.

For a fixed load and a given protocol $X = \{\text{FSA, FSA-RDP}\}$, we define $P_{L,X}^{opt}$ as the minimum $P_L$ obtained from Table 2 for different values of $V$, and denote by $V^*$ the value of $V$ at which $P_{L,X}^{opt}$ is achieved. For $V < V^*$, $P_{L,X} > P_{L,X}^{opt}$, as more collisions are expected due to the small number of data slots (FSA) or reservation slots (FSA-RDP). Then, packets stay longer in the buffer, which increases the probability that new packets are rejected (TD), or new packets push-out old packets from the buffer (PO). Also, for $V > V^*$, $P_{L,X} > P_{L,X}^{opt}$, the carried data packet rate does not grow substantially, but frame durations are longer (more reservation and data slots). As devices can only hold a single packet in their buffers, when the frame duration increases, multiple arrivals in a frame more likely occur, and losses increase accordingly.

For a given load, we define the inefficiency factor of protocol $X$ as $I_{f,X} = (P_{L,X} - P_{L,I})/P_{L,I}$. Note that $P_{L,I}$ corresponds to the packet loss probability achieved by FSA-RDP-I at the same load and protocol configuration (values of $V$ and $r$) at which $P_{L,X}^{opt}$ is found. Clearly, the smaller the value of $I_{f,X}$, the more efficient protocol $X$ is, as its $P_{L,X}$ is closer to the one achieved by the idealized protocol FSA-RDP-I ($P_{L,I}$). The inefficiency

<p>| Table 2: Minimum Data Packet Loss Probability ($P_L$) |
|---|---|---|</p>
<table>
<thead>
<tr>
<th>$V$</th>
<th>MAC</th>
<th>$r_T = 0.3$</th>
<th>$r_T = 0.5$</th>
<th>$r_T = 0.8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>FSA-RDP-I</td>
<td>0.0109</td>
<td>0.0338</td>
<td>0.1078</td>
<td></td>
</tr>
<tr>
<td>FSA-RDP</td>
<td>0.0151</td>
<td>0.0484</td>
<td>0.1539</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>$r_{opt}$</td>
<td>0.64</td>
<td>0.52</td>
<td>0.39</td>
</tr>
<tr>
<td>1</td>
<td>FSA</td>
<td>0.1343</td>
<td>0.3097</td>
<td>0.5228</td>
</tr>
<tr>
<td>2</td>
<td>$r_{opt}$</td>
<td>0.38</td>
<td>0.29</td>
<td>0.21</td>
</tr>
<tr>
<td>FSA-RDP-I</td>
<td>0.0115</td>
<td>0.0322</td>
<td>0.1020</td>
<td></td>
</tr>
<tr>
<td>sim</td>
<td>0.0115</td>
<td>0.0322</td>
<td>0.1021</td>
<td></td>
</tr>
<tr>
<td>FSA-RDP</td>
<td>0.0131</td>
<td>0.0411</td>
<td>0.1466</td>
<td></td>
</tr>
<tr>
<td>sim</td>
<td>0.0131</td>
<td>0.0411</td>
<td>0.1466</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>$r_{opt}$</td>
<td>1</td>
<td>1</td>
<td>0.79</td>
</tr>
<tr>
<td>3</td>
<td>FSA</td>
<td>0.1243</td>
<td>0.3036</td>
<td>0.5211</td>
</tr>
<tr>
<td>4</td>
<td>$r_{opt}$</td>
<td>0.76</td>
<td>0.56</td>
<td>0.41</td>
</tr>
<tr>
<td>FSA-RDP-I</td>
<td>0.0137</td>
<td>0.0354</td>
<td>0.1040</td>
<td></td>
</tr>
<tr>
<td>FSA-RDP</td>
<td>0.0154</td>
<td>0.0438</td>
<td>0.1413</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>$r_{opt}$</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>FSA</td>
<td>0.1194</td>
<td>0.2980</td>
<td>0.5196</td>
</tr>
<tr>
<td>6</td>
<td>$r_{opt}$</td>
<td>1</td>
<td>0.82</td>
<td>0.61</td>
</tr>
<tr>
<td>FSA-RDP-I</td>
<td>0.0162</td>
<td>0.0401</td>
<td>0.1103</td>
<td></td>
</tr>
<tr>
<td>sim</td>
<td>0.0162</td>
<td>0.0401</td>
<td>0.1103</td>
<td></td>
</tr>
<tr>
<td>FSA-RDP</td>
<td>0.0179</td>
<td>0.0486</td>
<td>0.1465</td>
<td></td>
</tr>
<tr>
<td>sim</td>
<td>0.0179</td>
<td>0.0486</td>
<td>0.1465</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>$r_{opt}$</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>FSA</td>
<td>0.1345</td>
<td>0.2945</td>
<td>0.5184</td>
</tr>
<tr>
<td>$r_{opt}$</td>
<td>1</td>
<td>0.82</td>
<td>0.61</td>
<td></td>
</tr>
<tr>
<td>FSA-RDP-I</td>
<td>0.0187</td>
<td>0.0451</td>
<td>0.1186</td>
<td></td>
</tr>
<tr>
<td>FSA-RDP</td>
<td>0.0205</td>
<td>0.0537</td>
<td>0.1542</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>$r_{opt}$</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>FSA</td>
<td>0.1534</td>
<td>0.3070</td>
<td>0.5173</td>
</tr>
<tr>
<td>$r_{opt}$</td>
<td>1</td>
<td>0.82</td>
<td>0.61</td>
<td></td>
</tr>
<tr>
<td>FSA-RDP-I</td>
<td>0.0212</td>
<td>0.0501</td>
<td>0.1277</td>
<td></td>
</tr>
<tr>
<td>FSA-RDP</td>
<td>0.0231</td>
<td>0.0590</td>
<td>0.1626</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>$r_{opt}$</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>FSA</td>
<td>0.1724</td>
<td>0.3255</td>
<td>0.5198</td>
</tr>
<tr>
<td>$r_{opt}$</td>
<td>1</td>
<td>0.82</td>
<td>0.61</td>
<td></td>
</tr>
</tbody>
</table>

<p>| Table 3: Protocol Inefficiency ($I_f$) |
|---|---|---|</p>
<table>
<thead>
<tr>
<th>MAC</th>
<th>$r_T = 0.3$</th>
<th>$r_T = 0.5$</th>
<th>$r_T = 0.8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>FSA-RDP</td>
<td>0.20</td>
<td>0.27</td>
<td>0.39</td>
</tr>
<tr>
<td>FSA</td>
<td>0.19</td>
<td>8.14</td>
<td>4.08</td>
</tr>
</tbody>
</table>
factors are shown in Table 3. Clearly, the efficiency improvement obtained by FSA-RDP respect to FSA is between one and two orders of magnitude, depending on the load. As observed, the difference between protocol efficiencies decreases as load increases. This effect was expected, as the contention increases with the load.

An interesting observation is that, for a fixed $V$, to achieve the minimum $P_L$ in FSA, the optimal permission probability, $r_{opt}$, has to be appropriately set (values lower than 1), particularly as load increases. However, for FSA-RDP, the optimal permission probability is constant to 1 for a large range of loads an values of $V$. This would make a scheme that adapts $r$ with the load more robust when deployed with FSA-RDP than with FSA.

6.2 Packet Delay Distributions

Figure 4 shows the CDF of packet delay for three representative offered loads $\rho_T = \{0.2, 0.5, 0.9\}$, $Q = 1$, and for $V = \{2, 4\}$. We consider the two queue management schemes, TD and PO. As a reference, we also display the ideal scheduling scheme, denoted by 'I'. For simplicity, we only consider the performance of the POI scheme, that outperforms the TDI scheme.

As expected, observe that packet delay is smaller for PO than for TD. As mentioned before, this is due to the fact that with PO, when a device buffer is full, eldest packets in the queue are pushed-out by the arrival of new packets. This helps to reduce the average packet delay. Observe also that the difference between packet delays obtained by the PO and TD schemes decreases as $V$ increases. This is due to the fact that as $V$ increases, a higher reservation success rate is obtained, the devices maintain their queue levels low, and the frequency of packet push-outs or tail drops decreases.

Figure 5 shows the 95th percentile of the packet delay with the offered load $\rho_T$. As expected, observe that packet delay is smaller for PO than for TD. As mentioned before, this is due to the fact that with PO, when a device buffer is full, eldest packets in the queue are pushed-out by the arrival of new packets. This helps to reduce the average packet delay. Observe also that the difference between packet delays obtained by the PO and TD schemes decreases as $V$ increases. This is due to the fact that as $V$ increases, a higher reservation success rate is obtained, the devices maintain their queue levels low, and the frequency of packet push-outs or tail drops decreases.

6.3 Devices with Larger Queues

In this section we extend the performance evaluation study to systems that deploy devices with buffers that can accommodate more than one data packet. This study is made by simulation. More precisely, we now deploy devices with two queue sizes, $Q = \{5, 10\}$ packets. We study two scenarios where the number of reservation mini-slots per RSF are $V = \{2, 4\}$, and the permission probability is always set to $r = 1$, unless otherwise stated. Here, we also have considered FIFO as the service discipline and two queue management schemes, TD and PO.

Figure 6 shows the evolution of the channel utilization $S$ with the load for systems with $V = 2, 4$. It is defined as,

$$S = W_{\gamma},$$

and expresses the fraction of time the channel is being occupied with data packets. That is, the RVP are considered control overhead. Although figures show results for the PO scheme, they coincide with the ones obtained for the TD scheme.

For $Q = \{1, 5, 10\}, V = 2$, the maximum achieved by FRA-RDP-I is $S_{max} = 0.908$. For $V = 4$ the curves are identical to the ones shown for $V = 2$. This similarity is expected. In the saturation regime all nodes are active in all frames, and as FRA-RDP-I is deployed, at each frame the number of data packets transmitted will be $V$ (provided that $M \geq V$). Then, the channel utilization in the saturation regime is given by

$$S_s = \frac{VW}{V + VW} = \frac{W}{1 + W},$$

i.e., $S_s$ is independent of $V$ and only depends on $W$. For $W = 10$, $M = 8$ and $V = 2$ or $V = 4$, we get $S_s = 10/11 = 0.909$, that approximately coincides with the maximum for FRA-RDP-I obtained by simulation and provided above. This explains why the curve of $S$ for $V = 2$ coincide with the one for $V = 4$ when FRA-RDP-I is deployed.
For $Q = 1$, $V = 2$, FRA-RDP peaks at $S_{\text{max}} = 0.698$. For $V = 2$, we observe that $S$ collapses for systems with $Q = 5$ for loads $\rho_T > 0.47$, and for systems with $Q = 10$ for loads $\rho_T > 0.40$. This phenomenon is due to the fact that, as load increases, devices get to a point where they stay active (with packets in their queues) during all system frames. That is, all devices contend for reservation slots in all frames. This leads to massive collisions, throughput reduction and massive losses.

This intuition can be validated using the analytical model and assuming that, in the saturation regime, $\pi_M = 1$ and $\pi_i = 0$, $i \neq M$. As an example, for $V = 2$, $r = 1$, the throughput in saturation (collapsed regime) can be computed from (4) simply by,

$$\gamma_s = \frac{\sum_{k=0}^{V} k f_k}{\sum_{k=0}^{V} t_k f_k}, \quad f_k = D^M_k (r = 1, V),$$

(29)

where $D^M_k (r = 1, V)$ is the probability of $k$ success reservations when $M$ devices contend, and was given by (2). Clearly, $S_s = W \gamma_s$. Then we get, $S_s (M = 8, V = 2, r = 1) = 0.238$, that coincides with the simulation value represented in Fig. 6a.

For $V = 4$, FRA-RDP achieves $S_{\text{max}} = 0.767$ when $Q = 5$, and $S_{\text{max}} = 0.756$ when $Q = 10$. In saturation they tend to $S_s = 0.721$. This is validated by computing analytically $S_s (M = 8, V = 4)$ using (29).

Figure 7 shows the evolution of the channel utilization in the saturation regime $S_s$ with the number of devices in the network $M$, when $r = 1$. The curve for
Fig. 7: Channel utilization as a function of the number of devices in the saturation regime.

Fig. 8: Fraction of cycles a device is inactive with the offered load $\rho_T$, $V = 2$.

$V = M$ displays $S_s$ for the optimal configuration of the RSF, i.e., the number of reservation mini-slots equals the number of contenders per frame. Recall that in the saturation regime all devices contend in all cycles. Curves for $V = M - 1$ and $V = M + 1$ display $S_s$ for suboptimal configurations of the RSF. These two curves are important as they show that as $M$ increases, the optimal configuration of the number of reservation mini-slots in the RSF becomes less critical. That is, the system performance achieved by an adaptive scheme that would configure $V$ as a function of the load would be almost insensitive to small errors in the estimation of $M$.

Figure 8 shows the evolution of the fraction of frames a device is inactive (empty queue) for a system with $V = 2$. As observed, when load grows above the values that lead to a system collapse in Fig. 6a, the corresponding devices become always active ($P_I = 0$), and therefore, all devices contend in all frames. This result helps to understand why the system with $Q = 10$ collapses earlier (smaller load) than with $Q = 5$.

Table 4: Representative values for the adaptive scheme

<table>
<thead>
<tr>
<th>$\rho_T$</th>
<th>0.2</th>
<th>0.4</th>
<th>0.6</th>
<th>0.7</th>
<th>0.75</th>
<th>0.8</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V$</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>$r$</td>
<td>1.00</td>
<td>0.95</td>
<td>0.65</td>
<td>0.55</td>
<td>0.50</td>
<td>0.50</td>
<td>0.40</td>
</tr>
</tbody>
</table>

Figures 9, 10 and 11 show the evolution of the packet loss probability $P_L$, the average packet delay $D$, and the 95th percentile of the packet delay $D_{95}$ with the total offered load $\rho_T$. The shape of the curves in these figures provide additional evidences of the system collapse phenomenon described previously. Note that the packet loss probability ($P_L$) curves for PO and TD coincide. For the average packet delay $D$, both schemes achieve similar performance at low loads. However, for high loads, the PO scheme achieves a slightly lower average packet delay than the TD scheme for the same load. This difference was explained before.

Observe that, the evolution of $D_{95}$ with the load has a characteristic knee shape. The knee occurs at approximately the same load for both the PO and the TD schemes, with negligible differences. Although not observable in Fig. 11a, $D_{95}$ for the PO scheme is slightly lower than the one for the TD scheme. The knee point occurs after the curves for $P_L$ start to grow exponentially. In the proximity of the knee, all devices are active in all frames, the throughput is very small as a lot of RVP collisions occur, and packets stay a long time in the queue.

A final observation is that as $V$ increases, the difference between the performance of FSA-RDP with $Q = 5$ and $Q = 10$ is drastically reduced, and they approach the performance of the FSA-RDP-I scheme. Note also that the performance of the FSA-RDP-I scheme for $V = 2$ and $V = 4$ is quite similar. This is due to the fact that the packet loss probabilities experienced by the ideal scheme for $V = 2$ and $V = 4$ are quite similar (in fact negligible) as shown in Fig. 9. This is an important observation, as it shows that their is still room for improvement in the design of new scheduling schemes that use more efficiently the system resources, particularly for small values of $V$.

6.4 Adaptive Scheme

One possible solution to the collapse condition problem displayed in Fig. 6a would be to deploy an adaptive scheme. It should adjusts both the number of reservation slots $V$ and the permission probability $r$ as a function of the load. Note that the system load varies with the product $M\lambda$. 
To validate this intuition we provide results for a system where devices have a buffer of \( Q = 10 \) packets. In this system, the controller adapts \( V \) and \( r \), and informs the devices by using, for example, the slot allocation packet (SAP) at the end of the RSF, or the ACK packet at the end of the DSF. The adaptation algorithm at the controller might be based on the number of successes and collisions observed during the last RSF. Then, when device \( i \) is active, it contends for one of the \( V \) reservation mini-slots with probability \( r \), where the values of \( V \) and \( r \) have been broadcast by the controller. Otherwise, with probability \( 1-r \), it sleeps until the next frame.

Table 4 shows some representative values for the adjusted \( V \) and \( r \) found by exhaustive search, when the initial value is \( V = 2 \) and for the load range \( \rho_T \in [0.2,0.9] \). Observe that at \( \rho_T = 0.75 \), \( V \) is increased by one unit.

The maximum channel utilization \( S_{\text{max}} \) can also be improved by adapting \( V \) and \( r \) as a function of the load. For example, Fig. 6a displays the result of adapting the system with \( Q = 10 \) in the interval \( \rho_T \in [0.2,1.6] \) (curve ‘PO adap’). \( S_s \) improves from 0.238 when a static configuration \( V = 2 \), \( r = 1 \), is used, to 0.8 when adaptation is used. Also, Fig. 6b displays the result of adapting the system with \( Q = 10 \) in the interval \( \rho_T \in [0.7,1.6] \) (curve ‘PO adap’). \( S_s \) improves from 0.721 when a static configuration \( V = 4 \), \( r = 1 \), is used, to 0.8 when adaptation is used. Note that as the load increases (\( \rho_T > 0.9 \)), \( V \) must be increased progressively up to \( V = 8 \).

Figures 9a, 10a, and 11a also show the impact that the adaptation of \( V \) and \( r \) have on the system performance. As observed, the system performance improves drastically. \( P_L \), \( D \), \( D_90 \) all decrease. This allows the system to successfully operate over a wider load range than without adaptation. Figure 8 also shows the beneficial impact that adaptation has on \( P_T \).

From the energy efficiency point of view, it is desirable to avoid collisions in the reservation phase. That is, it is desirable to achieve an energy consumption per successfully transmitted bit as low as possible. Focusing on the system with \( Q = 10 \) as an example, it is clear that when the load goes beyond the collapse point, the system remains in a state where the energy is wasted by the transmission of RVP that collide, and a very low throughput is obtained, achieving a poor energy efficiency. Then, an adaptive scheme would drastically help to reduce energy consumption, and improve energy efficiency.

7 Conclusions

In this paper we studied a type of Frame Slotted ALOHA protocol that deploys reservation and data packets, that we refer to as FSA-RDP. In FSA-RDP frames are divided in two subframes, the reservation and the data subframes. In reservation subframes active devices that have access permission contend for sending reservation packets to the controller. Those that succeed (their reservation packets do not collide) are assigned a dedicated data slot to transfer a data packet during the data subframe.

We model the operation of the FSA-RDP protocol with a discrete-time Markov chain, and determine the protocol efficiency and the packet delay distribution, for a scenario where M2M devices have a buffer to store a single packet. We define the protocol efficiency in terms of the data packet loss probability. We compare the efficiency of the FSA-RDP protocol and the conventional Frame Slotted ALOHA (FSA), with that of an idealistic version of the FSA-RDP protocol (FSA-RDP-I), which efficiency can be considered as an upper bound for the FSA-RDP protocol. Results show that the protocol efficiency of FSA-RDP is between one and two orders of magnitude larger than the efficiency of the conventional FSA. In addition, the efficiency of FSA-RDP is close to the one obtained for FSA-RDP-I when the number of reservation mini-slots and the permission probability are properly configured.

We determine the cumulative distribution function of the packet delay for FSA-RDP, when both the tail drop (TD) and push-out (PO) queue management schemes are deployed. We study different loads and configurations for the protocol. In the scenarios studied, results show that the packet delay for PO is smaller than the one obtained for TD. However, the difference between the delay obtained by both queue management schemes is only significant for high loads, and when the number of mini-slots in the reservation subframe is small. That is, when the collision rate is high.

We also studied the system performance with buffer sizes of 5 and 10 packets. We observed that, as the load increases, the system can collapse, bringing the throughput to very low values and, correspondingly, the packet loss probability and packet delay to a very high ones. An adaptive scheme that adjust the number of reservation slots \( V \) and the permission probability \( r \) at every frame is suggested to cope with this problem. To validate this suggestion, the performance of the system is measured when the adaptation is performed manually by exhaustive search of the optimum parameter values. Results show a substantial performance improvement when adaptation is performed.
Fig. 9: Packet throughput with the offered load $\rho_T$.

Fig. 10: Average packet delay with the offered load $\rho_T$.

Future work will be oriented to the design of efficient adaptive schemes. In addition, a testbed implementation will be considered to validate experimentally the conclusions drawn from the analytical and simulation models.

Acknowledgements

This work has been supported by the Ministry of Economy and Competitiveness of Spain through projects TIN2013-47272-C2-1-R and TEC2015-71932-REDT. The authors would like to thank the support received from the Instituto ITACA (Instituto Universitario de Tecnologías de la Información y Comunicaciones) at the Universitat Politècnica de València, Spain.

References

1. 3GPP, TS 22.368. Service Requirements for Machine-Type Communications, 2004
packets for M2M networks. In: 5th Int. Conf. on Future Internet of Things and Cloud (FiCloud), pp. 18–24. IEEE (2017)