Medium Access Control Techniques for Massive Machine-Type Communications in Cellular IoT Networks

Carlos A. Astudillo[1]* **and Nelson L.S. da Fonseca**[2]
[1]São Paulo State University, Institute of Geosciences and Exact Sciences
Rio Claro, Brazil, 13506-900
[2]State University of Campinas, Institute of Computing
Campinas, Brazil, 13083-852
carlos.astudillo@unesp.br, nfonseca@ic.unicamp.br

Abstract

A key component of the Internet of things (IoT) ecosystem is wide-area network connectivity, for which cellular network technologies are a promising option through their support of massive machine-type communications (mMTC). However, numerous devices transmitting sporadically small data packets in a highly synchronized way can generate overload on the radio access network. This situation leads to a shortage of resources, especially those associated with the random access procedure for contention, control, and data transmission, causing *preamble collision*, *control message blocking*, and *data collision*. As a result, mMTC traffic can jeopardize the provisioning of quality of service (QoS) to end-devices, decrease the network efficiency, and increase access latency and device energy consumption. This paper proposes medium access control (MAC) techniques for addressing problems related to the support of mMTC in cellular networks. First, a solution for allocating control resources with QoS provisioning and access differentiation is provided, including a resource management model, a scheduling algorithm, and a message prioritization policy. Second, a mechanism for reducing *data collision*s is also proposed. This solution comprises a protocol in which every device with scheduled retransmission uses a probabilistic policy to decide whether to retransmit, a novel method at the device to estimate the number of nodes trying random access, and two different retransmission policies employing this estimation. Results show that the proposals support QoS, decrease access latency, decrease the device energy consumption, and increase resource utilization under massive random access.

Keywords: Cellular Networks, Internet of Things, Massive-Machine Type Communications, Medium Access Control.

1 Introduction

The Internet of things (IoT) is one of the major technological trends transforming our society by improving quality of life, increasing industry productivity, and creating new business opportunities [1]. It involves numerous devices equipped with sensing, computing, and communication capabilities, such as sensors, actuators, machines, and vehicles [2]. Several industry sectors benefit from the information exchange in the IoT, such as transportation, health care, manufacturing, agriculture, smart cities, smart grid, and smart home.

The third generation partnership project (3GPP) cellular network technologies are gaining momentum in the low power wide area (LPWA) IoT connectivity landscape [3] due to their capacity to overcome issues of the unlicensed such as interference, reliability, and availability, coverage, ubiquity, and management. However, conventional human-type communication (HTC)-oriented technologies have several limitations when dealing with machine-type communication (MTC) traffic [4]. The 3GPP has made many efforts to improve the support of IoT applications in cellular networks, which resulted in the specification of the cellular Internet of things (CIoT)

*This work was developed while the author was with the Institute of Computing, State University of Campinas.

technologies, including enhancements for MTC support in existing cellular technologies, *e.g.*, LTE-MTC (LTE-M) and narrowband IoT (NB-IoT) [5].

The fifth generation (5G) networks cover three broad use case families: enhanced mobile broadband (eMBB), massive machine-type communications (mMTC) and ultra-reliable and low-latency communications (URLLC). To meet their stringent requirements, the 5G standard encompasses both the evolution of the long term evolution (LTE) technology (*e.g.*, the LTE enhancements for MTC such as LTE-M and NB-IoT) and the addition of a new radio access technology known as new radio (NR) [6]. Since NB-IoT and LTE-M fulfill the 5G network requirements for mMTC services, they both were recognized as IMT-2020 5G standards and will evolve as part of the 5G specifications in the 3GPP [5].

The IoT is exponentially increasing the number of devices connected to the Internet as never seen before. However, this huge number of expected MTC devices transmitting sporadically small packets in a highly synchronized way puts very high pressure on the aforementioned cellular network technologies for mMTC. This traffic pattern leads to shortage of radio resources, especially those associated with the random access protocol, which is typically used by the IoT devices to request resources for sporadic uplink packet transmissions.

The *shortage of random access resources* rises naturally with massive random access attempts since many devices enter the random access procedure simultaneously, a situation quite common in massive Internet of things (MIoT) scenarios. It is the root cause of various problems introduced in cellular networks due to the support of mMTC services, such as *preamble collision*, *control message blocking*, and *data collision*.

The *preamble collision* occurs when two or more devices choose the same preamble in a given random access subframe. Despite the name of this problem, a preamble under collision is typically successfully detected by the base station [7]. Thus, the base station physical (PHY) entity passes the list of detected preambles including non-collided and collided preambles to its medium access control (MAC) entity for subsequent resource allocation [7]. This is due to the physical characteristic of the preamble signals and their detection technique, typically based on energy peak searching in the physical random access channel (PRACH) power delay profile (PDP) [8]. Although this problem can occur even with HTC traffic, its impact is much more notorious with mMTC traffic due to the high probability of several devices simultaneously transmitting random access preambles.

Since the amount of network resources is limited, the *control message blocking* can also occur in the random access procedure. This problem happens because resources available to respond to the successfully detected preambles (MSG1) within a random access response window and to successfully decoded data transmissions (MSG3) before the expiration of the contention resolution timer may not be enough. In such a situation, blocking of control messages due to resource limitations may cause that some devices with a successfully detected preamble or successfully received data message do not receive the random access response (RAR) message and the contention resolution (CR) message to continue/finish the random access procedure. In both cases, the device has to perform a new RACH trial after a backoff time, further increasing the random access channel (RACH) load. The number of RAR (equivalently, uplink grants) and CR messages that a base station can deliver in time is limited and depends on the available downlink control resources as well as on uplink data resources. Moreover, the problem can be still worst in networks with coexistence of MTC devices and HTC users. In such a scenario, devices with different quality of service (QoS) requirements and scheduling mechanisms (*e.g.*, dynamic scheduling, semi-persistent scheduling, and random access) compete for the shared network resources.

Once the base station allocates radio resources to detected preambles by means of a RAR messages including an uplink grant, all devices involved in a *preamble collision* receive the same allocation. Consequently, their data transmissions (MSG3) collide with high probability, generating a *data collision*. A *data collision* also involves *collision of data retransmission*s because the hybrid automatic repeat request (HARQ) protocol is employed in the random access procedure for protecting the data transmission from channel impairments. Thus, both the random access schemes and the resources allocation mechanisms are important to be addressed to improve the support of mMTC in cellular networks.

This paper aims *at proposing and evaluating medium access control techniques for improving the support of massive machine-type communications in 3GPP cellular networks*. It investigates the *shortage of random access resources* and the problems derived from it. Specifically, mechanisms to deal with the effects of the *control message blocking* on the QoS provisioning and to alleviate the problems of *collision of data retransmission*s are provided. This paper focuses on the interplay between random access and resource allocation to improve the support of massive machine-type communications in 3GPP cellular networks.

The main set of contributions of this paper can be stated as follows:

C1. **A random access resource allocation scheme for supporting mMTC in cellular networks:** It includes a control resource management model and control scheduling algorithms for access differentiation in the random access (RA) procedure. This contribution is an extension of our work published in [9–12]. This paper, however, provides details on the control resource management model, system model, and proposed scheduling policies.

C2. **A probabilistic data retransmission protocol for the random access procedure:** The proposed protocol deals with the *problem of collision during data retransmission* in the random access procedure. It exploits probability theory to improve the performance of the random access procedure by increasing the efficiency of its *data transmission phase* under preamble collision, quite common in massive Internet of things scenarios. This contribution extends the work in [13] by adding the derivation and analysis of the probability distribution of contending MTC devices per detected preamble, justifying the approach followed in the proposal of the protocol.

The rest of this paper is organized as follows. Table 1 presents the list of acronyms used in the paper. Section 2 reviews the relevant related work. Section 3 presents the system model used and assumpitons made in this work. Section 4 presents the packet downlink control channel (PDCCH) resource management model employed in this work. Section 5 describes the proposed scheme for scheduling random access control messages, whereas Section 6 introduces the proposed probabilistic retransmissions approach. Section 7 describes the performance evaluation methodology employed, and Section 8 shows the simulation results and discusses the performance of the proposed solutions. Finally, Section 9 concludes the paper.

2 Related Work

On the one hand, the random access procedure in cellular networks involves access differentiation (*e.g.*, prioritization or isolation) among different devices or traffic types. The random access resources are, however, inherently shared by all devices performing contantion-based and non-contantion based random access in the cell, as explained in Section 3.3.1. Besides the natural isolation and prioritization intended by the non-contention based random access mode, some random access schemes also provide means for access differentiation by splitting the preambles for contention-based random access into two or more groups.

These groups can be created either dividing the preambles in the code-domain as in the RACH resource separation (RRS)-based schemes [4, 14, 15], or in the power-domain as in [16, 17].

RRS allows preamble sequence separation between HTC and MTC in order to alleviate the effect of the MTC on the HTC [4]. By reserving an exclusive set of preamble sequences for a small number of user equipments, the collision probability in the preamble transmission phase of the random access procedure is significantly reduced in the presence of massive access attempts. Condoluci *et al.* [18] also introduced a random access scheme that reserves a set of preamble sequences for transmitting critical alarm messages.

Kim *et al.* [16] proposed Prioritized Random Access (PRA) scheme that enables fixed-location MTC devices to indicate their priority (either low or high) during the random access procedure by means of the transmit power level of the preamble sequence. In [17], the same authors have extended their idea to coexisting MTC/HTC scenarios, in which HTC users get high priority and MTC devices low priority. The proposed scheme includes a procedure to create first the MSG2 messages for high-priority preambles and then those for low-priority preambles. In this way, the assembled MSG2 messages to be scheduled contain the time advancing command of the high priority users, increasing the chances of successful transmission of MSG3 messages from high priority users since they use the time advancing matching technique to reduce collisions in MSG3 transmissions. However, the actual scheduling of PDCCH resources for this messages is neglected and the authors assume that unlimited PDCCH resources are available. Moreover, any PDCCH allocation framework nor PDCCH scheduler is described in their work. In addition, since this procedure is made in a TTI-basis, MSG2 messages arriving in the n-th random access opportunity (RAO) are positioned in the queue after low priority MSG2 messages from a previous RAO, which can impact the high priority users when PRACH configurations with various RAO per frame are used.

On the other hand, the random access schemes for machine-to-machine communications have recently attracted a lot of attention in the research community because they have an important role in the support of IoT over cellular networks. Two main techniques can be used [19]: the contention-avoidance schemes, which aim at reducing the number of attempts under high loads, impacting the preamble transmission phase of the random access procedure, and the contention-resolution schemes, which aim at resolving the collisions among MTC devices during the random access procedure. In the former, the access class barring (ACB) and extended access barring (EAB) are the main approaches [4], both barring some devices to attempt random access during high random access loads. However, in the following, we focus on the review of mechanisms that reduce MSG3 collisions, including contention-resolution solutions.

Table 1: List of acronyms

2G second generation.	**PDU** packet data unit.
3GPP third generation partnership project.	**PHICH** physical HARQ indicator channel.
5G fifth generation.	**PHY** physical.
ACB access class barring.	**PMF** probability mass function.
ACK acknowledgement.	**PPA** preamble-priority-aware.
BCH broadcast channel.	**PRACH** physical random access channel.
BLER block error rate.	**PRB** physical resource block.
BSR buffer status report.	**PS** packet scheduling.
C-RNTI cell radio network temporary identifier.	**PUCCH** physical uplink control channel.
CAM critical alarm message.	**PUSCH** packet uplink shared channel.
CBR constant bit rate.	**QoS** quality of service.
CBRA contention-based random access.	**RA** random access.
CCE control channel element.	**RA-RNTI** random access radio network temporary identifier.
CFRA non-contention based random access.	**RACH** random access channel.
CIoT cellular Internet of things.	**RAN** radio access network.
CR contention resolution.	**RAO** random access opportunity.
CSS common search space.	**RAP** random access prioritized.
DCI downlink control information.	**RAPID** random access preamble identifier.
DL downlink.	**RAR** random access response.
DL-SCH downlink shared channel.	**RAT** radio access technology.
EAB extended access barring.	**RB** resource block.
EDT early data transmission.	**RLC** radio link control.
eMBB enhanced mobile broadband.	**RRC** radio resource control.
FD frequency-domain.	**RRM** radio resource management.
FDD frequency division duplexing.	**RRS** RACH resource separation.
HARQ hybrid automatic repeat request.	**RSRP** received signal reference power.
HTC human-type communication.	**RV** random variable.
IoT Internet of things.	**SAP** service access point.
LPWA low power wide area.	**SC-FDMA** single-carrier frequency division multiple access.
LTE long term evolution.	**SI** system information.
LTE-M LTE-MTC.	**SIC** self-interference cancellation.
LTE-Sim LTE simulator.	**SID** silent descriptor.
MAC medium access control.	**SNR** signal-to-noise ratio.
MCS modulation and coding scheme.	**SPS** semi-persistent scheduling.
MIoT massive Internet of things.	**SR** scheduling request.
mMTC massive machine-type communications.	**TA** timing advance.
MNO mobile network operator.	**TC-RNTI** temporary cell radio network temporary identifier.
MTC machine-type communication.	**TD** time-domain.
NACK negative-acknowledgment.	**TDPS** time-domain packet scheduling.
NB-IoT narrowband IoT.	**TPC** transmit power control.
NOMA non-orthogonal multiple access.	**TTI** transmission time interval.
NORA non-orthogonal random access.	**UE** user equipment.
NR new radio.	**UL** uplink.
PDCCH packet downlink control channel.	**UL-SCH** uplink shared channel.
pdf probability density function.	**URLLC** ultra-reliable and low-latency communications.
PDP power delay profile.	**VAD** voice activity detection.
PDSCH physical downlink shared channel.	**VoIP** voice over IP

Ali *et al.* [19] proposed a contention resolution based on an m-ary contention tree splitting technique. The scheme is based on a non-standard random access technique that allows transmitting the preamble sequence jointly with the user equipment (UE) identity, which is used to identify collided preambles. A binary tree is created for each collided preamble and the RAR message is used to inform the collided UE about the resource to be used in the next random access attempt, sending uplink grants to uncollided preamble only. The process is finalized when all collisions are resolved. Vilgelm *et al.* [20] [21] propose a random access protocol for LTE networks based on binary countdown technique for contention resolution. This protocol introduces micro-slots before MSG3 transmission for prioritizing MTC devices and resolving contention. This approach also assumes that MTC devices can listen to the transmission of each other. In general, the main problem with the contention-resolution-based random access approaches is that extra signaling between the base station and the user equipments or among user equipments is required to introduce the solution into the LTE protocol, making them difficult to be implemented in commercial cellular networks.

Kim *et al.* [22] proposed a random access procedure in which multiple RAR messages per detected preamble are sent in order to reduce the number of MSG3 collisions due to preamble collision. However, this scheme has some drawbacks. The base station does not have information about the number of MTC devices per collided preamble. So, it allocates multiple uplink grants per detected preamble based on the estimated expected value of that variable. However, this may waste a lot of physical uplink control channel (PUCCH) resources, which decreases the resource availability for actual user data. Magrin *et al.* [23] introduced a method to estimate the number of user equipments that chose the same preamble based on machine learning techniques. Even though the authors showed promising results with a synthetically generated dataset, the proposed technique adds additional complexity to the base station, and the collection of a real dataset is difficult with current base station implementations. Thus, the proposed technique is not easily implementable in existing cellular networks.

The capture effect, which allows the decoding of one of the interfering signals, was also exploited in random access schemes for increasing the MSG3 detection probability by applying power ramping technique [24] or multiple power levels [25] to the MSG3 transmissions. However, these approaches increase the MTC device energy consumption [24] as well as the interference that the packet uplink shared channel (PUSCH) can cause to PUSCH/PRACH of neighboring cells in co-channel deployments.

Ko *et al.* [26] introduced a mechanism that avoids sending multiple MSG3 messages based on the time-advanced commands received as part of the RAR message. However, this approach can be exploited just in MTC devices with no mobility and when the collided MTC devices are located at different distance from the base station. Even if the capture effect could be exploited, it allows for the decoding of a maximum of one user transmission per RAR message.

Liang *et al.* [27] proposed the non-orthogonal random access (NORA) scheme, employing self-interference cancellation (SIC). This scheme introduced a technique to detect preamble collisions and exploits the use of power-domain non-orthogonal multiple access (NOMA) to decode more than one MSG3 messages per detected preamble. The main limitations of this proposal include the reduced chance of detecting a preamble collision in small cells as well as the increase in base station complexity due to the SIC receiver and the superimposed preamble detection.

In summary, existing control scheduling approaches do not support random access schemes in their intended goal of providing differentiated random access. Even though the above-mentioned random access schemes, as well as others in the literature, can provide differentiation among preamble groups in the *signature transmission* phase, current algorithms for scheduling of control messages do not consider the random access priority associated to individual random access-related control messages of the subsequent phases of the random access procedure. Moreover, contention-avoidance random access schemes just avoid some MTC devices to send preamble, whereas contention-resolution random access schemes effectively tackle the *MSG3 collision problem* by avoiding devices to transmit their MSG3 messages. However, existing techniques to address the *MSG3 collision problem* requires additional signaling messages and are based on nonstandard compliant procedures that make them difficult to be implemented in commercial networks. This paper aims to address these two gaps identified in the literature.

3 System Model and Assumptions

This section presents the system model and assumptions used in this work. Section 3.1 presents the network layout and network-level settings. The physical layer configuration is described in Section 3.2, while the relevant MAC layer aspects are detailed in Section 3.3. Finally, the traffic models are presented in Section 3.4. The list of symbols used in the paper is provided in Table 2.

Table 2: List of symbols

		$\mathcal{D}$	set of PDCCH message classes.
A	total number of orthogonal preambles available in the cell.	$\mathcal{P}$	set of priority levels for PDCCH message requests.
B	cell bandwith.	$\mathcal{RB}$	set of available resource blocks.
I	number of idle preambles in a random access subframe.	$\mathcal{R}$	set of DCI message requests to schedule on the PDCCH.
K	number of contending MTC devices.	$\mathcal{Z}$	set of RAR messages to schedule.
L_{CBR}	packet size of CBR traffic.	Q_{PDCCH}	priority queue of DCI message requests to schedule on the PDCCH.
L_{IoT}	packet size of IoT traffic.		
$N_{\text{CCEs}}^{\text{PDCCH}}$	number of CCEs for the PDCCH.	Q_{RAR}	priority queue of RAR messages.
N_{CCEs}	total number of CCEs for downlink control channels.	ϕ	fraction of control resources assigned to the PDCCH.
N_{RAR}	maximum number of grants per random access response MAC packet data unit.	r_{HTC}	number of contention-based random access preambles for HTC users.
$N_{\text{RBs}}^{\text{PRACH}}$	number of resource blocks for the PRACH.	r_{MTC}	number of contention-based random access preambles for MTC devices.
$N_{\text{RBs}}^{\text{PUCCH}}$	number of resource blocks for control signaling on the PUCCH.	b	number of PDCCH message classes.
$N_{\text{RBs}}^{\text{PUSCH}}$	number of resource blocks for uplink data transmission on the PUSCH.	d	number of preambles detected in a random access subframe.
N_{RBs}	total number of resource blocks in the cell.	i	random access subframe index.
$N_{\text{UEs}}^{\text{HTC}}$	number of HTC users.	j	preamble index.
$N_{\text{UEs}}^{\text{MTC}}$	number of MTC devices.	l	preamble transmission attempt index.
O	number of devices in collision with a collided MTC device.	m	number of DCI message requests to schedule on the PDCCH.
P_C	preamble collision probability.	o	number of devices in collision with a collided MTC device.
P_S	random access success probability.		
P_{RTX}	retransmission probability given a preamble collision event.	p_{ACB}^{*}	optimal ACB access probability.
		p_e	MSG3 transmission error probability.
P_{TX}	HARQ transmission probability.	p_{ACB}	ACB access probability.
Q	maximum number of MTC devices that may be in a collision with a collided device.	p	number of priority levels for control messages.
T_{CBR}	packet-arrival periodicity of CBR traffic.	r	number of preambles available for contention-based RA.
T_{IoT}	packet-arrival periodicity of IoT traffic.		
T_{dist}	distribution time.	u	MSG3 transmission attempt index.
W	number of transmitting MTC devices per detected preamble.	x	instance of ON period.
		y	instance of OFF period.
γ	cell radius.	z	number of RAR messages to schedule.

3.1 Network Layout and Network-level Assumptions

A traditional hexagonal cellular network layout [28] is considered, as illustrated in Figure 1. It is made of six neighbour cells and a target cell, which is served by a base station supporting both HTC and MTC devices. In what follows, we focus on the description of the target cell, including its base station and devices within its coverage. The cell coverage is approximated as a circular cell shape with γ m radius. The cell bandwidth is B MHz in the frequency division duplexing (FDD) mode. There are N_{RBs} physical resource blocks available in each direction, following the 3GPP standards for the given bandwidth and technology [29]. At the center of the cell, a single base station serves several devices uniformly located around it. There are $N_{\text{UEs}}^{\text{HTC}}$ HTC users and $N_{\text{UEs}}^{\text{MTC}}$ MTC devices.

MTC devices are assumed to be in the radio resource control (RRC) idle or inactive states. HTC users are considered to be in the RRC connected state. There are two different types of HTC users in the cell: *attached* and *handover* users. The former group is made of users that are currently registered and attached to the target cell base station, whereas the latter group is composed of those users that are within the coverage of the target cell performing handover from neighbour cell base stations (where they are attached) to the target cell base station. Devices performing random access are assumed to have successfully decoded the synchronization signals (PSS and SSS) and acquired relevant system information such as the PRACH configuration parameters.

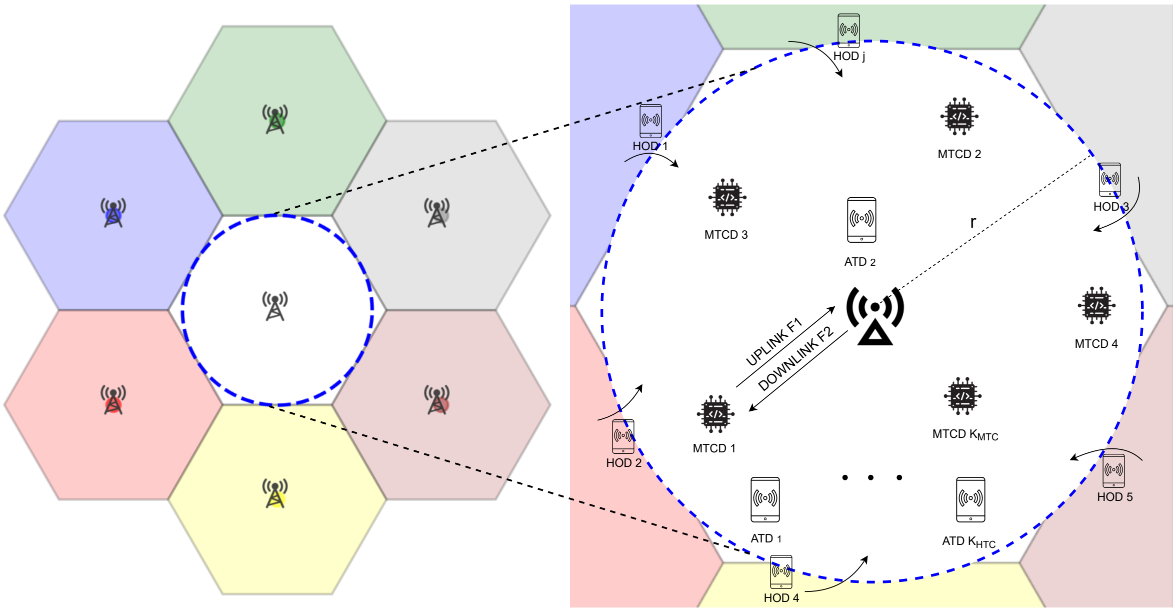


Figure 1: Hexagonal network layout. MTCD, ATD, and HOD stand for MTC device, *attached* HTC device, and *handover* HTC device, respectively.

3.2 PHY Layer

The downlink control channels are configured to use a total of N_{CCEs} control channel elements. The number of CCEs for the PDCCH $\left(N_{\text{CCEs}}^{\text{PDCCH}}\right)$ is assumed to be a fraction ϕ of the total number of CCEs for downlink control channels such that $N_{\text{CCEs}}^{\text{PDCCH}} = \phi \cdot N_{\text{CCEs}}$ [30]. The physical downlink shared channel (PDSCH) is assumed to occupy the rest of the downlink capacity.

In the uplink, the PUSCH consists of $N_{\text{RBs}}^{\text{PUSCH}}$ physical resource blocks, used for uplink user data and control message transmissions, with the remaining $N_{\text{RBs}}^{\text{PUCCH}}$ physical resource blocks reserved for control signaling transmission on the PUCCH such that $N_{\text{RBs}}^{\text{PUSCH}} + N_{\text{RBs}}^{\text{PUCCH}} = N_{\text{RBs}}$. The physical resource blocks used by PUCCH are always located at the beginning and at the end of the spectrum to avoid fragmentation in the scheduling of PUSCH resources due to the contiguity constraint of the single-carrier frequency division multiple access (SC-FDMA) operation. In a RACH occasion, however, the PRACH supports A orthogonal preambles and uses the central $N_{\text{RBs}}^{\text{PRACH}}$ physical resource blocks such that $N_{\text{RBs}}^{\text{PRACH}} + N_{\text{RBs}}^{\text{PUSCH}} + N_{\text{RBs}}^{\text{PUCCH}} = N_{\text{RBs}}$.

Uplink data transmissions on the PUSCH passes through a channel model considering four physical phenomena: path loss, penetration loss, shadowing, and multi path [31]. Moreover, Gaussian noise is also included in the signal-to-noise ratio (SNR) values for PUSCH transmissions. The probability of successful PUSCH transmission is given by mapping the SNR value for the transmission and the block error rate (BLER) curve for the modulation and coding scheme (MCS) used in the transmission.

This model also integrates the preamble transmission and reception process, including the power control mechanism for the uplink transmissions with power ramping technique, and the detection of the preamble based on the preamble received power. The uplink transmit power control in radio access technologies is a key radio resource management function, providing adequate link quality to transmit while minimizing the energy consumption of the battery-constrained wireless devices and the interference to other users of the system. The PRACH transmit power (P_{PRACH}) for a given preamble transmission is defined in [32, 33] and given by:

$$P_{PRACH} = \min\{P_{UE}, P_{TARGET} + PL\}; \tag{1}$$

P_{UE} is the maximum UE transmit power, PL is the path-loss factor estimated by the device, and

$$P_{TARGET} = P_{INITIAL} + \Delta P_{TYPE} + (I-1) \cdot i, \tag{2}$$

where $P_{INITIAL}$ is the expected power to be received at the base station, ΔP_{TYPE} is a constant associated with the preamble type defined for the cell, I is the index of this preamble transmission, and i is the power ramping step value [34, 35].

The parameters used by the device to perform the random access procedure, including the above-mentioned $P_{INITIAL}$ and S for PRACH power control, are sent by the base station in the SIB2 [34, 35], while the PL is locally estimated by device calculating the difference between the received signal reference power (RSRP) and the base station reference signal transmission power in the downlink (value also included in the SIB2).

The transmissions of control signaling in both downlink and uplink, however, are assumed to occur without errors due to perfect link adaptation and fixed and low modulation and coding scheme employed by them. This assumption is widely used when assessing the random access and packet scheduling performance and holds for transmissions of the MSG2 and MSG4 on the PDSCH, scheduling requests transmissions on the PUCCH and downlink control information message transmissions (used for signaling control messages and data transmissions) on the PDCCH.

3.3 MAC Layer

The non-contention based random access mode is employed by the *handover* users, while the contention-based random access mode is used by the *attached* HTC users and MTC devices. The non-contention based random access (CFRA) and contention-based random access (CBRA) were modelled as described in Section 3.3.1.

Besides the conventional CBRA scheme, the RRS scheme is also supported in the network. In the RRS scheme, the r preambles for contention-based random access are further divided into those for HTC devices and those for MTC devices such that $r = r_{\text{HTC}} + r_{\text{MTC}}$. The RRS scheme was proposed to reduce the impact of mMTC on the random access performance of traditional HTC users.

3.3.1 The Random Access Procedure

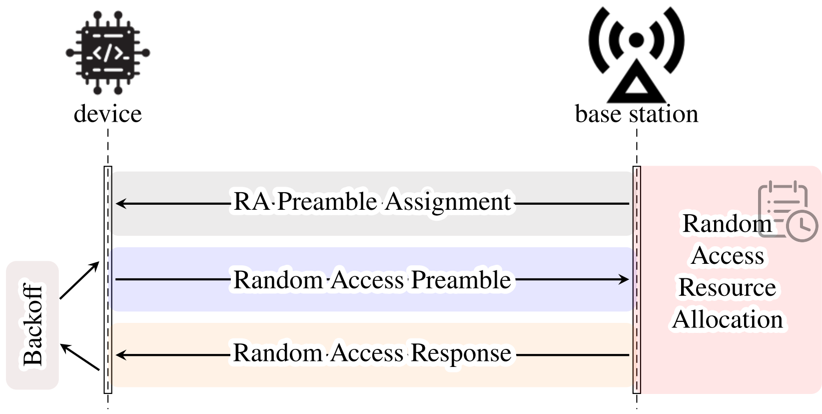
The random access procedure model employed in 3GPP radio access network technologies is illustrated in Figure 2 and works as follows. It can be executed in two operational modes: non-contention (Figure 2a) and contention based (Figure 2b). The former is used to perform handover or to re-establish synchronization prior to downlink data transmission, while the latter is used in the following cases: (*i*) initial access to the network, *i.e.*, when the radio interface is turned on or after a long period of network inactivity; (*ii*) to request uplink resources upon arrival of uplink packets at the device buffer if data and control resources are not assigned to the device; (*iii*) to re-establish connection after a radio failure; (*iv*) loss of uplink synchronization; and (*v*) transition from inactive to connected states.

In both modes, a device transmits an orthogonal random access preamble on the random access channel (MSG1). In the contention-based mode, the preamble is randomly selected by the device from the set of contention-based preambles, which is periodically updated by the base station through the system information sends on the broadcast channel (BCH). Conversely, in the non-contention mode, the preamble is explicitly assigned by the base station via downlink signaling. In this mode, the preamble comes from a set of unique preambles dedicated exclusively for this mode to avoid random access *preamble collision*. The sum of the these two disjoint sets gives the total number of orthogonal preambles available in the cell.

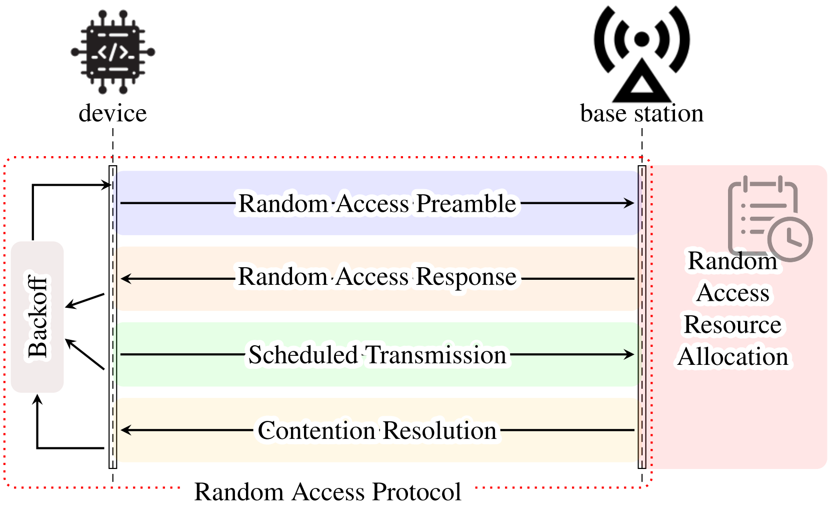
The base station performs preamble detection and timing advance (TA) estimation on the PRACH signals received at every RACH occasion, hereinafter called a random access subframe. The list of preambles detected and their associated timing information is then passed to the MAC layer, which is responsible for allocating uplink resources to the detected preambles. The base station executes a resource allocation algorithm and informs the devices about its allocation through a RAR message addressed to the random access preamble identifier (RAPID) - MSG2, which is a control message sent on the downlink shared channel (DL-SCH) containing an uplink grant for data transmission on the uplink shared channel (UL-SCH) and a TA command for uplink synchronization.

If at least one RAR message needs to be scheduled in a given subframe, a downlink control information message addressed to the random access radio network temporary identifier (RA-RNTI) must be scheduled on the common search space (CSS) region of the PDCCH to inform devices about the PDSCH resources on which the RAR message(s) are transmitted. The RA-RNTI uniquely identifies the time-frequency resources on which the preamble detected was transmitted. For instance, two preambles (different or not) transmitted on the same random access subframe and PRACH have the same RA-RNTI.

Once a RAR message addressed to its transmitted RAPID is decoded by the device, it transmits a scheduled message on the UL-SCH (MSG3) following the information contained in the RAR message. The MSG3 can be a connection setup/resume request. If the early data transmission (EDT) [36] feature is enabled, actual user data can also be transmitted. The MSG3 transmission supports the HARQ technique, used to protect scheduled data transmissions from channel impairments and hardware imperfections. Thus, if an MSG3 transmission is



(a) The non-contention based mode.



(b) The four-step contention-based mode.

Figure 2: The 3GPP random access procedure.

unsuccessful, its retransmission can be performed. In the contention-free mode, the random access procedure is finished once the device receives the RAR message.

Finally, upon successful reception of the MSG3, the base station sends a CR message on the DL-SCH (MSG4) to the device. If this message contains the device identity, the contention-based random access procedure is finished. Note that the number of control messages that a base station can issue in a given subframe is limited and depends on the data and control resources available in the cell.

If no RAR message within a RAR window is found, a CR message is not received before the contention resolution timer expires, or the maximum number of MSG3 HARQ transmissions is achieved, the device transmits a new random access preamble after a random backoff period provided that the maximum number of preamble transmissions has not been achieved.

3.3.2 Device MAC Entity

Both padding and regular buffer status reports (BSRs) are configured to be sent by HTC devices in the RRC connected state. For devices with dynamic scheduling, one byte is used to convey a padding BSR message. A regular BSR is triggered when data arrives in an empty radio link control (RLC) buffer at the device. When a device does not have uplink resources for transmission of a regular MAC BSR control element on the PUSCH, the device triggers a scheduling request transmission. When a device needs to send a scheduling request message and there is no available PUCCH resources for it, the device initializes the random access procedure.

3.4 Application Layer

This section describes the application traffic supported by the network devices. Each device is assumed to use a single type of traffic to avoid interference with intra-device MAC scheduling. MTC devices transmit IoT traffic, whereas HTC users can transmit either voice over IP (VoIP), video or constant bit rate (CBR) traffic.

3.4.1 Device Activation

The *attached* HTC users are assumed to be transmitting/receiving data. Both semi-persistent and dynamic scheduling are supported by the network. The *attached* HTC users employing dynamic scheduling are assumed to have PUCCH resources allocated to send scheduling request messages so that they do not frequently use the random access procedure to request uplink resources. The *attached* HTC users employing the random access procedure are assumed to try their first random access attempt (initial activation) following a Poisson distribution with rate λ_{HTC} within a given time period [28, 37]. Thus, the probability that there are n device activations in a time period is given by

$$\Phi_{\lambda_{HTC}}(n) = \frac{\lambda_{HTC}^n}{n!} e^{-\lambda_{HTC}}. \tag{3}$$

The handover requests on the RACH of *handover* HTC users from neighbour cells are assumed to arrive to the target cell following also a Poisson distribution with rate λ_{HO} [28, 37]. Similar to (3), we have:

$$\Phi_{\lambda_{HO}}(n) = \frac{\lambda_{HO}^n}{n!} e^{-\lambda_{HO}}. \tag{4}$$

Activation of MTC devices considers a *bursty traffic scenario*, which simulates MTC device transmissions highly synchronized in a cell. This scenario simulates the response to an emergency event, such as earthquake alarm or fire alarms, as suggested by the 3GPP in [4]. The MTC device arrivals (activation of the MTC device in the scenario) follows a time-limited Beta probability density function $p(\tau)$ within a period interval $T_{\text{dist}} = 10\,\text{s}$ [4], given by

$$p(\tau) = \frac{\tau^{\alpha-1}(T-\tau)^{\beta-1}}{T^{\alpha+\beta-2}\text{Beta}(\alpha,\beta)}, \tag{5}$$

where Beta(α, β) is the Beta function with the constant parameters $\alpha = 3$ and $\beta = 4$.

3.4.2 IoT Traffic

The MTC devices are assumed to have IoT traffic. This traffic is modeled as a single data packet that fits the smallest transport-block size available in the radio access technology considered. This means that the device needs either a single resource block (RB) to transmit the packet or that it fits into the transmission capacity of the early data transmission technique, in which the data packet is carried by the MSG3. The corresponding transmission procedure is triggered with the activation of the device in the scenario as described above. After the activation, the MTC devices are assumed to generate an IoT packet of L_{IoT} bytes every T_{IoT} s.

3.4.3 HTC Traffic

Each VoIP source is simulated as a G.729 voice flow modeled as an ON/OFF system, *i.e.*, a Markov chain with two states, as illustrated in Figure 3. The random variable of the ON periods (X) are exponentially distributed with mean λ, *i.e.*,

$$\mathbb{P}(x) = \lambda e^{-\lambda x}, \tag{6}$$

and the random variable of the OFF periods (Y) have a truncated exponential probability density function (pdf) with an upper limit of T_{max} and an average value $T_{OFF} = \left({}^{1}\!/_{\mu}\right)$

$$\mathbb{P}(y) = \frac{\lambda e^{-\lambda y}}{1 - e^{-\lambda y}}, \tag{7}$$

During the ON period, a VoIP packet of 20 bytes arrives every 20 ms (*i.e.*, the source data rate is 8 kb/s), while during the OFF period, voice activity detection (VAD) is assumed and just a silent descriptor (SID) packet every $160\,\text{ms}$ is generated.

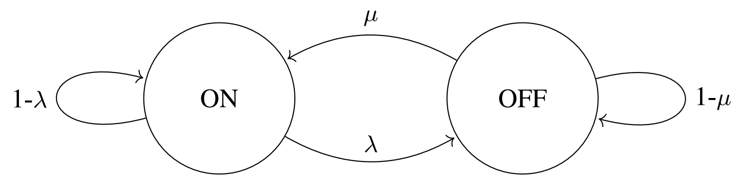


Figure 3: ON/OFF Markov model for VoIP sources.

VoIP traffic uses semi-persistent uplink scheduling [38], so that resources on the PUSCH are allocated periodically, without the need for the device to send scheduling request (SR) request and for the base station to send uplink grants on the PDCCH. Uplink resources for data transmission on the PUSCH are reserved every 20 ms after the first VoIP packet transmission.

However, to decrease the amount of PDCCH resources used by the VoIP users and to accommodate more HTC users per cell [39, 40], semi-persistent scheduling (SPS) with initial random access mechanism [41] is assumed to be used by VoIP users.

CBR sources produce a packet of L_{CBR} bytes every T_{CBR} s, while video sources are based on real traffic traces.

4 The PDCCH Resource Management Model

This section presents the proposed scheme to manage random access-related messages into the PDCCH manager introduced in [42], and it explains how these messages are scheduled.

Figure 4 gives an overview of the base station MAC entity and the relevant information required for resource allocation in the random access procedure. This section closes the gap in the literature about a holistic view of the base station MAC entity considering the random access procedure and the control and data plane scheduling. Particularly, it focuses on the interaction of the PDCCH manager with the random access and packet schedulers, which are other important MAC functions at the base station.

Entities illustrated in blue in Figure 4 perform a specific radio resource management (RRM) function or control function at the MAC layer, while those in orange represent information assumed to be available at the base station MAC layer for proper RRM operation. In this model, packet scheduling (PS) is implemented in two (decoupled) steps, one in the time-domain and the other in the frequency-domain, which significantly reduces the complexity of the scheduling process [43]. The support of QoS requirements is controlled by the time-domain scheduler, leaving the frequency-domain scheduler to perform radio channel-aware scheduling.

4.1 The Random Access Manager

The random access manager has two main components, the random access control and the random access scheduler. The random access control handles the random access procedure at the MAC layer and the random access scheduler is in charge of processing the random access-related control messages such as random access response, contention resolution messages, and scheduling request messages. The random access scheduler allocates PDSCH resources for transmitting MSG2 and MSG4 messages, and PUSCH resources for responding to detected preambles and scheduling request sent in the MSG3 messages. The random access manager also assembles the content of the RAR messages inside the MAC packet data units and interacts with the PDCCH manager for asking for downlink control information messages for signaling the corresponding PDSCH and PUSCH allocations.

4.2 The PDCCH Manager

For every downlink subframe, the PDCCH manager determines the downlink control information messages, their aggregation level (the number of control channel elements (CCEs) used to transmit the message), and the physical PDCCH resources used to transmit the scheduled messages. The limited number of the CCEs available in a subframe and the level of aggregation used by the downlink control information messages can have a large effect on network performance, especially in MIoT scenarios. These aspects define how many downlink control information (DCI) messages can be transmitted by a given base station [44].

The PDCCH manager works in three steps, as shown in Figure 5. One to prioritize the requests, another to perform the message scheduling and resource reservation, and the other to assemble and deliver the downlink

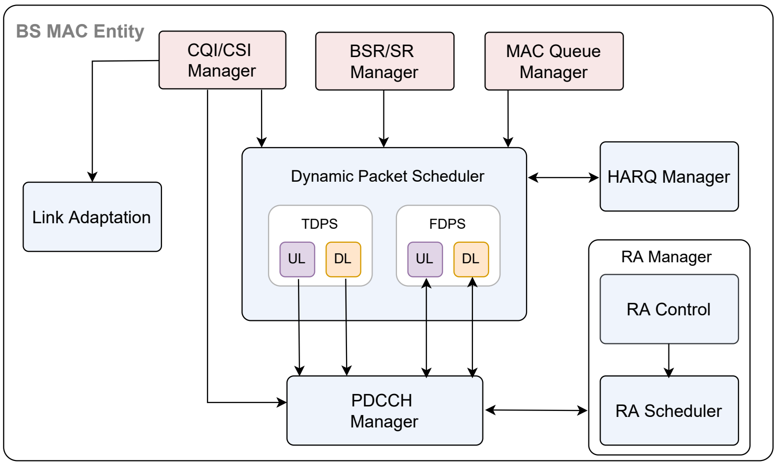


Figure 4: Resource allocation functional architecture.

control information packet data units.

4.2.1 PDCCH Message Request Prioritization

In the first step, Request Prioritization, a priority list is created with all PDCCH messages to be scheduled by the PDCCH manager in the current scheduling round. This priority list must consider not only uplink grants and downlink assignments, but also all other control messages that must be scheduled on the PDCCH, such as random-access related messages (*i.e.*, random access response and CR messages), BSR grants, and transmit power control (TPC) commands.

The input to this task is a list of PDCCH message requests, as follows: (*i*) **uplink grants** and **downlink assignments** for specific devices, which are passed by the uplink time-domain packet scheduling (TDPS) and downlink TDPS, respectively; (*ii*) **RAR MAC PDU message** and **contention resolution messages** passed by the random access manager. These requests require *downlink assignments* for signaling their transmission on the PDSCH; (*iii*) **BSR grants** passed also by the random access manager to grant uplink resources to respond to a scheduling request performed through the random access procedure by a device in RRC-connected state; (*iv*) **TPC commands** passed by the MAC control to ask some device in the RRC-connected state to adjust its transmit power.

The downlink control information messages for requests in item (*i*) are addressed to the device cell radio network temporary identifier (C-RNTI) and CR messages in item (*ii*) needs to be addressed to the temporary cell radio network temporary identifier (TC-RNTI), while downlink control information messages for random access response requests are addressed to the RA-RNTI. The downlink control information messages of items (*iii*) and (*iv*) are addressed to the C-RNTI.

The priority queue is created by executing a *PDCCH scheduling policy*.

4.2.2 Actual Resource Allocation

The second step involves the assignment of the messages to a given number of CCEs (link adaptation) as well as assignment of a power offset (power control), so that the block error rate target can be met.

In addition, in this step, the PDCCH manager performs the scheduling of the CCEs. The CCEs on the PDCCH are reserved as a function of the necessity of each message (*i.e.*, aggregation level, search space and transmit power). Resources are assigned first to the head of the priority queue; if no resource is available for that, the CCE scheduler moves to the next message in the queue. The messages in the priority queue are scheduled either sequentially, or in a way to find the best allocation of resources for each message independently of ordering. Here, the former approach, respecting the priorities of the messages in the queue, has been adopted.

Figure 6 illustrates the main steps of the basic RRM algorithm for the PDCCH employing the RRM model described in this section.

After each message is processed, the PDCCH manager provides the queues to the frequency-domain packet schedulings as well as information on which devices can be scheduled in the uplink and downlink directions. The

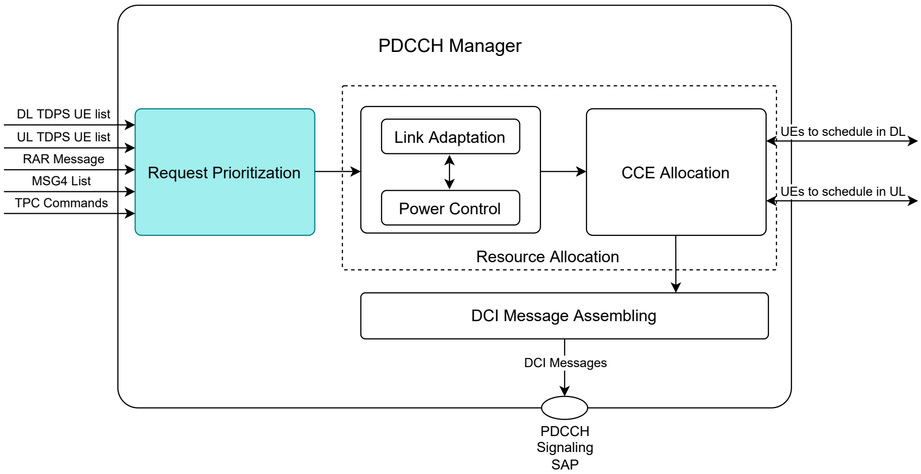


Figure 5: Functional structure of the PDCCH manager at the base station MAC entity.

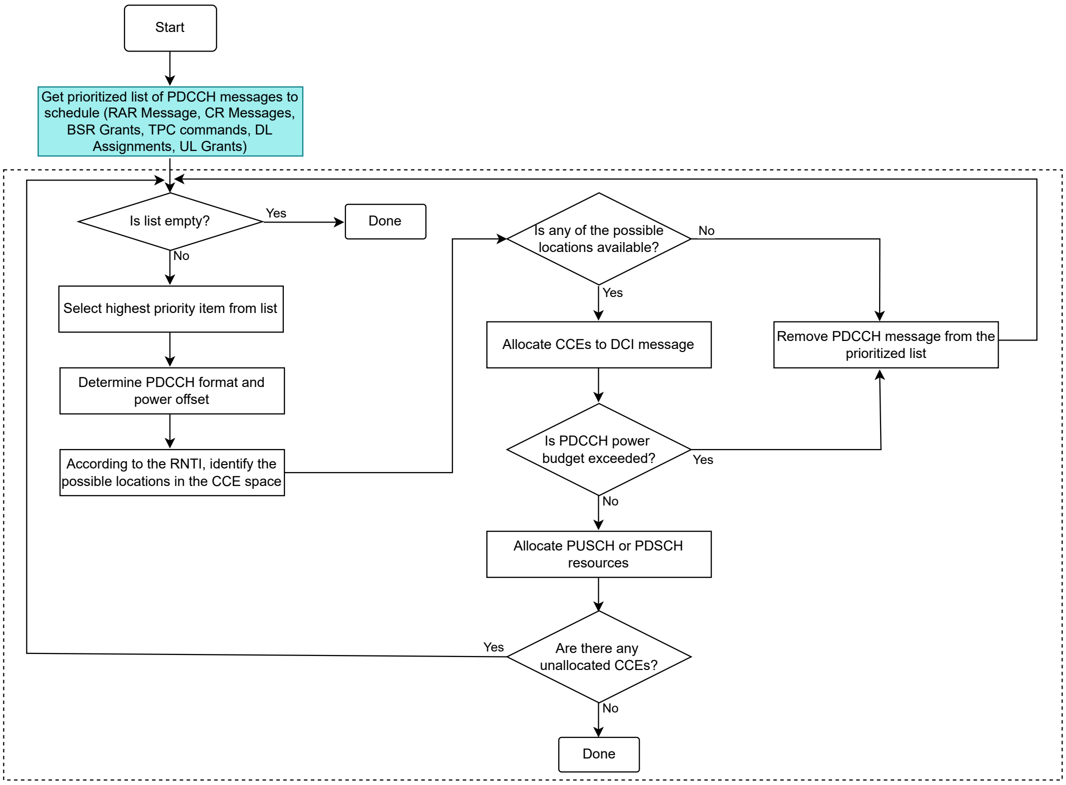


Figure 6: PDCCH scheduling process (adapted from [45] to a fixed number of control channel element).

frequency-domain packet schedulings, then, allocate the physical resource blocks (PRBs) on the uplink/downlink shared data channels.

4.2.3 Downlink Control Information Message Assembling

In this step, the PDCCH manager uses information on the DCI message requests scheduled and devices selected to assemble DCI messages for each uplink grant and downlink assignment and transmits them on the PDCCH. Moreover, the random access related messages can have their DCI messages assembled immediately after the CCEs allocation task finishes. After a DCI message is assembled, the PDCCH manager sends it to the physical layer entity through the PDCCH signaling service access point (SAP) to deliver it to the devices.

4.3 The Random Access Prioritized PDCCH Scheduling Policy

In this scheduling policy, decreasing levels of priority are given to the following messages: RAR message, CR messages, BSR grants, TPC commands, DL assignments and UL grants. To be fair between the uplink and downlink performance, downlink assignments and uplink grants are interleaved as long as there are both type of messages to schedule so that one uplink grant is prioritized for each downlink assignment included. In this policy, the order of the priority of the uplink grants and downlink assignments received from the time-domain packet scheduling are maintained by the PDCCH scheduler in the assignments of priorities.

5 The Preamble-Priority-Aware Approach for Scheduling Random Access Control Messages

This section introduces the preamble-priority-aware (PPA) approach for scheduling random access control messages. The main objective of the PPA scheme is to provide the base station with the ability to support random access schemes in their purpose of providing differentiated network access in the random access procedure under heavy RACH loads. Thus, the proposal deals with the treatment of random access response (MSG2), contention resolution (MSG4) messages in the entire resource allocation process, including their scheduling and corresponding allocation signaling.

The PPA solution involves three main components: (*i*) preamble priority definition, which is discussed in Section 5.1; (*ii*) random access scheduling, which is described in Section 5.2; and (*iii*) PDCCH scheduling, which is explained in Section 5.3. Moreover, the complexity analysis of the PPA scheme is provided in Section 5.4.

5.1 Definition of the Preamble Priorities

This section answers the following question: *how one can establish the priority of different preamble groups?*

Since the random access differentiation is performed in the very first step of the random access procedure (the *signature transmission* phase) by the existing random access schemes, the preamble priority levels rely on the underline scheme used in the network. The preamble-priority-aware scheme defines first the priority between preambles for contention-based random access and non-contention based random access, and then it considers the priorities in each of the groups.

No existing resource allocation algorithm for control scheduling in the literature considers preamble priority in contemporary cellular networks. In second generation (2G) networks, high priority values had been given to handover calls over new calls since blocking of a handover call is less desirable than blocking of a new call [28, 46]. This is equivalent to give high priority to non-contention based random access over contention-based random access in contemporary cellular systems. The rationality is that a handover request (performed using non-contention based random access) comes from a device that is already attached to the system with a cell context and typically has an ongoing active session, for example a video or voice call. Therefore, preambles for non-contention based random access receives the highest priority level in the preamble-priority-aware scheme.

Once the priority between non-contention and contention-based random access modes are established, the priority levels of the preambles for contention-based random access in the preamble-priority-aware approach are considered. The preamble-priority levels among contention-based preambles depend on the intended treatment given to them by the underline random access scheme, which includes differentiation strategies for HTC/MTC coexistence, emergency alarms, performance isolation, and low-latency access.

Since the traditional 3GPP contention-based random access scheme does not provide any differentiation among contention-based preambles[1], the base station does not have the means to differentiate subsequent control messages in the random access procedure. Thus, when the traditional random access scheme is used, all contention-based preambles are mapped onto the lowest priority level. Medium- and low-priority levels are, however, defined for those random access schemes that differentiate between two contention-based preamble groups, such as in the RACH resource separation schemes [4], the preamble-power based random access schemes [16, 17], and Condoluci *et al*'s proposal [14]. For instance, the RACH resource separation scheme [4] divides contention-based preambles into MTC and HTC groups, while the Condoluci *et al*'s mechanism uses an exclusive group of contention-based preambles to transmit very high priority critical alarm message [14]. In these cases, the preamble-priority-aware approach maps the set of high priority preambles and low priority preambles into the

[1]Even though groups A and B of contention-based preambles are supported by 3GPP standards in the traditional random access procedure [32, 33], they are intended to provide a mean of indicating the size of the MSG3 grant.

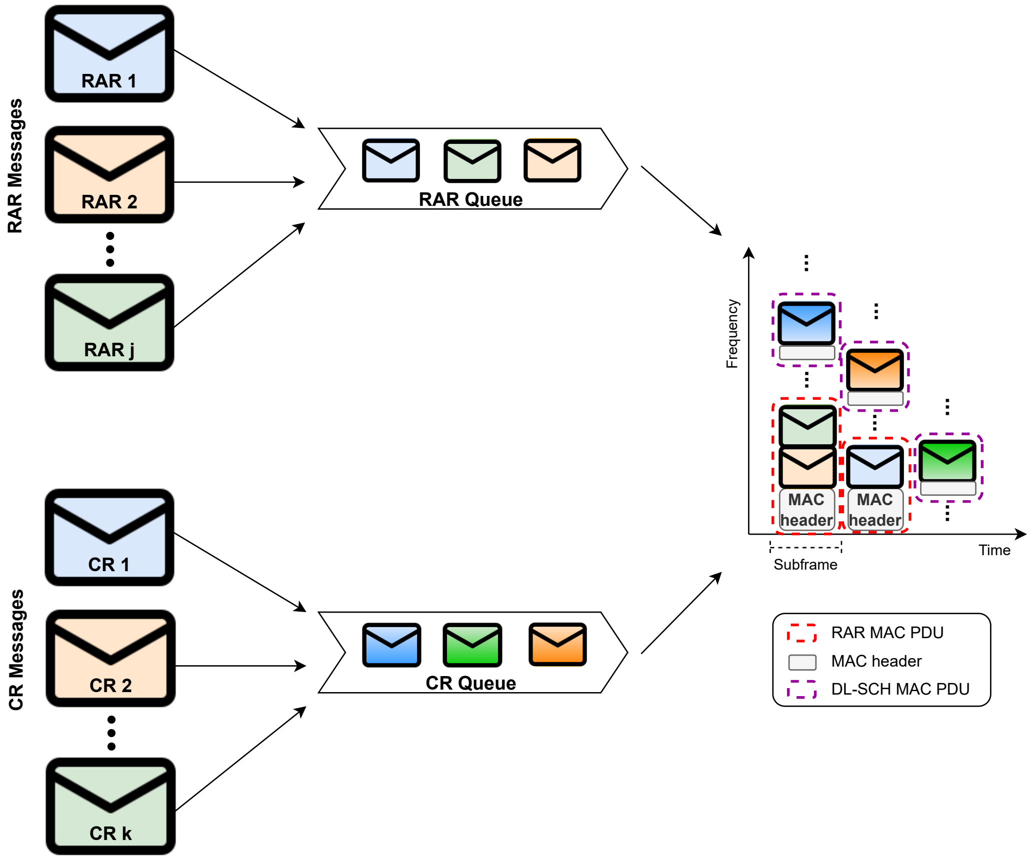


Figure 7: Functional illustration of the random access manager on the downlink data channel. DL-SCH: downlink shared channel; RAR: random access response; CR: contention resolution; PDU: packet data unit

medium-priority and low-priority levels, respectively. More priority levels can exist depending on the random access scheme used.

5.2 The Preamble-Priority-Aware Random Access Scheduling Policy

Once the preamble priorities are defined, this section discusses the functionality of the random access scheduler and its interaction with the PDCCH manager described in Section 4 into the context of the preamble-priority-aware scheme.

The random access scheduler is to the random access control messages what the packet scheduler is to the user data packets. It allocates resources on the PDSCH for the transmission of random access response and contention resolution messages from the base station to the devices. The random access manager maintains two queues, one with the random access response messages to schedule (RAR queue) and another with the contention resolution messages (CR queue), as illustrated in Figure 7. Similar to resource allocation for data packet transmission on the downlink shared channel, the control message transmissions also have to be signaled on the PDCCH through downlink control information messages. Thus, at each PDCCH scheduling round, the random access scheduler sends a list of PDCCH message requests to the PDCCH manager. The list contains requests for signaling transmissions of (*i*) contention resolution (MSG4) messages enqueued at the base station and (*ii*) at most a random access response MAC packet data unit, which can carry up to N_{RAR} random access response (MSG2) messages. To decrease the computational complexity, a reduced number of requests for contention resolution messages can be passed to the PDCCH manager.

In the preamble-priority-aware scheme, the MSG4 prioritization task is left to the PDCCH manager, while the prioritization of RAR messages to be assembled into a random access response MAC packet data unit is performed by the random access scheduler. This approach is useful to provide quality of service in the PDCCH resource

Algorithm 1 The preamble-priority-aware random access scheduling policy.

Require:

$\mathcal{Z}$ ▷ set of RAR messages to schedule
$\mathcal{RB}$ ▷ set of available resource blocks
$\mathcal{D}$ ▷ set of PDCCH message classes

Ensure:

Q_{RAR} ▷ priority queue of RAR messages

1: **for each** $z \in \mathcal{Z}$ **do**
2:     **for each** $d \in \mathcal{D}$ **do**
3:         **if** z.getPriorityLevel() $==$ d.priorityLevel() **then**
4:             $p = d$.getPriority()
5:             Q_{RAR}.enqueue(z, p)
6: $q \leftarrow 0$
7: **while** $q < N_{\text{RAR}}$ & Q_{RAR}.size() > 0 & $\mathcal{RB}$.isRbAvailable() **do**
8:     $m_q \leftarrow Q_{RAR}$.dequeue()
9:     m_q.setUlGrant($\mathcal{RB}$.reservedEdgeRB())
10:     Add m_q to the random access response MAC packet data unit
11:     $q \leftarrow q + 1$

allocation process. The random access scheduler maintains the prioritization of the random access response messages assembled into a random access response MAC packet data unit, while its transmission signaling is scheduled by the PDCCH manager.

Once the PDCCH manager schedules the downlink control information message to transmit a random access response on the downlink shared channel, the random access scheduler selects the random access response messages that are actually assembled into the scheduled random access response MAC packet data unit and performs allocation of uplink shared channel resources to be included into the uplink grants that are sent into each random access response message scheduled. The physical resource blocks allocated for MSG3 transmissions are those at the edges of the available spectrum to avoid fragmentation due to contiguity single-carrier frequency division multiple access constraint.

The preamble-priority-aware random access scheduling policy is described in Algorithm 1. It serves the random access response messages in order of priority, as discussed in Section 5.1. First, the priority of each random access response message to schedule is obtained and then the message is added to the queue (Lines 1 to 5). The preamble-priority-aware approach first assembles random access response messages to respond to highest-priority level preambles (those for non-contention based random access). Afterward, it schedules random access response messages for contention-based random access from medium-priority level, if applicable. Finally, random access response messages to low-priority levels are served, as defined in Section 5.1. This process is executed until all random access response messages are served or the maximum number of grants per random access response MAC packet data unit are achieved (Lines 7 to 11).

5.3 The Preamble-Priority-Aware PDCCH Scheduling Policy

In the PDCCH manager, the preamble-priority-aware scheme must follow the same approach as in the random access prioritized and MSG2-first PDCCH scheduling policies to guarantee that the random access-related control messages received adequate treatment in the PDCCH resource allocation process. Thus, the preamble-priority-aware PDCCH scheduling policy first serves the PDCCH random access response message request, if any, and then it serves the contention resolution message requests. However, the contention resolution message requests are served in decreasing order of priority of the preamble sequences that generated them, as defined in Section 5.1.

Once all PDCCH requests for MSG4 messages are scheduled, the other control messages are processed as in the random access prioritized algorithm. The preamble-priority-aware scheduling policy is illustrated in Algorithm 2. The algorithm iterates over all the request (Lines 2 to 4), determining the priority of the request (Lines 6 to 9). After this, each request is added to the prioritized queue of PDCCH requests that will be served by employing the algorithm in Figure 6.

Algorithm 2 The preamble-priority-aware PDCCH scheduling policy.

Require:
$\mathcal{R}$ ▷ set of DCI message requests to schedule on the PDCCH
$\mathcal{P}$ ▷ set of priority levels for PDCCH message requests
Ensure:
Q_{PDCCH} ▷ priority queue of DCI message requests to schedule on the PDCCH
$Q_{PDCCH} \leftarrow \emptyset$ ▷ Initialize the priority queue
for each $r \in \mathcal{R}$ **do**
    pv = getRequestPriority(r)
    Q_{PDCCH}.enqueue(r, pv)
function GETREQUESTPRIORITY(r)
    for each $p \in \mathcal{P}$ **do**
        if r.getType() == p.getTypeName() **then**
            return p.getPriorityValue()
    return 0

5.4 Complexity Analysis

This section analyzes the time complexity of the preamble-priority-aware proposal, which involves two different components of the base station MAC entity, the PDCCH manager and the random access scheduler.

The preamble-priority-aware PDCCH scheduling policy first performs the classification of the requests, which takes $\mathcal{O}(m \times (b + p - 1))$, where m, b, and p are, respectively, the number of DCI message requests to schedule on the PDCCH, number of PDCCH message classes, and number of priority levels for control messages. After this, adding the requests to the priority queue according to their priority takes $\mathcal{O}(m)$. Thus, the complexity is $\mathcal{O}(m \times (b + p - 1) + m)$. Since b and p are small constants, the time complexity of the preamble-priority-aware PDCCH scheduling policy is $\mathcal{O}(m)$.

The part corresponding to the random access scheduler is analyzed as follows. Let z be the number of RAR messages to schedule in a subframe. The random access scheduler performs RAR message prioritization, which takes $\mathcal{O}(z \times p)$. Then, the random access scheduler adds the RAR messages to the MSG2 priority queue, which takes $\mathcal{O}(z)$. Thus, the complexity is $\mathcal{O}(z \times p + z)$. Since p is a small constant, the time complexity of the PPA random access scheduler is $\mathcal{O}(z)$. Thus, the final complexity of the PPA approach is $\mathcal{O}(m + z)$.

6 Probabilistic Retransmissions for the Random Access Procedure

This section introduces the probabilistic retransmission approach proposed for the random access procedure. Section 6.1 derives and analyzes the distribution and expectation of the number of MTC devices transmitting a detected preamble in a random access subframe (W), which justifies the approach followed in the proposed protocol. Section 6.2 describes the design of the retransmission protocol, while Section 6.3 proposes two retransmission policies for use by the devices with the proposed protocol. Section 6.4 describes a lightweight method for estimating the network load in terms of the number of contending devices in a given RACH occasion, which is required by the proposed policies. Section 6.5 introduces a novel technique based on RAR message counting to estimate the average number of detected preambles at the device, a variable that is not available at the device and that is required in the network load estimation.

6.1 Number of Transmitting Devices per Detected Preamble

In this section, the distribution and expectation of the number of MTC devices transmitting a detected preamble in a random access subframe is derived and analyzed.

Theorem 6.1 (Probability distribution and expectation of W)**.** *If $K \geq 1$ devices are trying random access simultaneously employing the conventional random access scheme with r preambles available for contention-based random access, then the probability mass function of the number of MTC devices transmitting a detected preamble in a random access subframe (W) is given by*

$$\mathbb{P}_W(x) = \begin{cases} \mathbb{P}\left(W = x\right) & \text{, if } 1 \leq x \leq K \\ 0 & \text{, otherwise} \end{cases}, \tag{8}$$

where

$$\mathbb{P}\left(W=x\right)=\frac{\binom{K}{x}\left(\frac{1}{r}\right)^{x}\left(1-\frac{1}{r}\right)^{K-x}}{1-\left(1-\frac{1}{r}\right)^{K}}\,. \tag{9}$$

Moreover, the expected value of W is given by

$$\mathbb{E}\left[W\right]=\frac{K}{r\left[1-\left(1-\frac{1}{r}\right)^{K}\right]}\,, \tag{10}$$

Proof. Assume that each of K devices trying random access randomly selects a preamble in a random access subframe. The base station detects a set of preambles at ith random access subframe ($\mathcal{Y}_i$).

Let $\mathbf{X_i}=X_{i1},X_{i2},\cdots,X_{ir}$ be a random vector, where X_{ij} denote the random variable of the number of devices transmitting the jth preamble in the ith random access subframe. The random vector $\mathbf{X_i}$ has a multinomial distribution with index K and parameter vector $\mathbf{p}={}^1\!/r,\forall j\}$. Note that a preamble j is *detected* at the ith random access subframe (*i.e.*, $j\in\mathcal{Y}_i$) if and only if $X_{ij}\geq 1$.

Since $p_j={}^1\!/r,\forall j$, $\mathbf{X_i}=\mathbf{X}=X_1,X_2,\cdots,X_r\ \forall i$, $X_j=X\ \forall j$, and X follows a binomial distribution with parameters $p=p_j={}^1\!/r$ and K, given by

$$\mathbb{P}\left(X=x\right)=\mathbb{B}_x(K,p)=\binom{K}{x}p^{x}\left(1-p\right)^{K-x}\,. \tag{11}$$

Let A be the event that a preamble is *detected* in a random access subframe. The probability of A, denoted by $\mathbb{P}\left(A\right)=\mathbb{P}(X\geq 1)$ is given by

$$\mathbb{P}\left(A\right)=\mathbb{P}(X\geq 1)=1-\mathbb{B}_0(K,p)=1-\left(1-\frac{1}{r}\right)^{K}\,. \tag{12}$$

Let W_d denote the random variable of the number of devices transmitting dth detected preamble in a random access subframe. The probability mass function of W_d can be calculated based on (11) and (12) as

$$\begin{aligned}\mathbb{P}\left(W_d=x\right)&=\mathbb{P}\left(X=x|A\right),1\leq x\leq K\\&=\frac{\mathbb{P}\left(X=x\right)}{\mathbb{P}\left(A\right)}\\&=\frac{\mathbb{B}_x(K,p)}{1-\mathbb{B}_0(K,p)}\\&=\frac{\binom{K}{x}p^{x}\left(1-p\right)^{K-x}}{1-\left(1-p\right)^{K}}\,.\end{aligned} \tag{13}$$

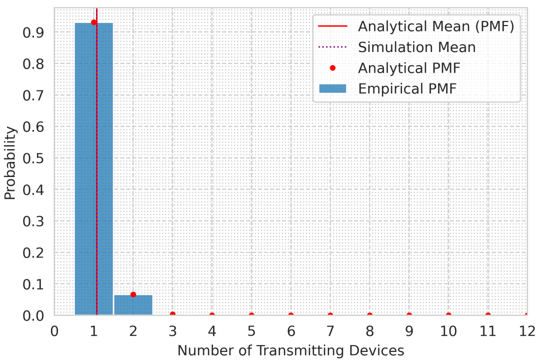
The second part of the theorem can be proved by employing the probability mass function (PMF) just derived as follows.

$$\begin{aligned}\mathbb{E}\left[W_d\right]&=\sum_{x=1}^{K}x\cdot\mathbb{P}\left(W_d=x\right)\\&=\sum_{x=1}^{K}x\cdot\frac{\binom{K}{x}p^{x}\left(1-p\right)^{K-x}}{1-\left(1-p\right)^{K}}\\&=\frac{1}{1-\left(1-p\right)^{K}}\cdot\sum_{x=1}^{K}x\binom{K}{x}p^{x}\left(1-p\right)^{K-x}\\&=\frac{K\cdot p}{1-\left(1-p\right)^{K}}\,,\end{aligned} \tag{14}$$

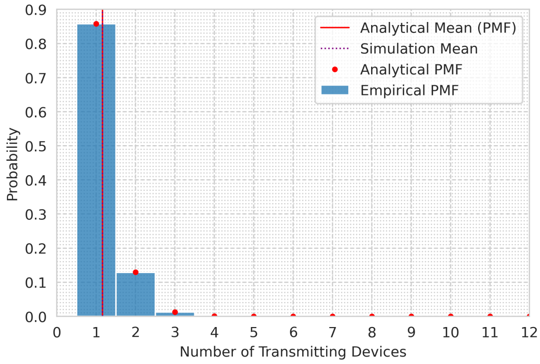
where the last step is obtained by the expectation of the binomial distribution.

After substituting p by ${}^1\!/r$ in (13) and (14), the probability mass function and expected value of W in Theorem 6.1 follow, respectively.

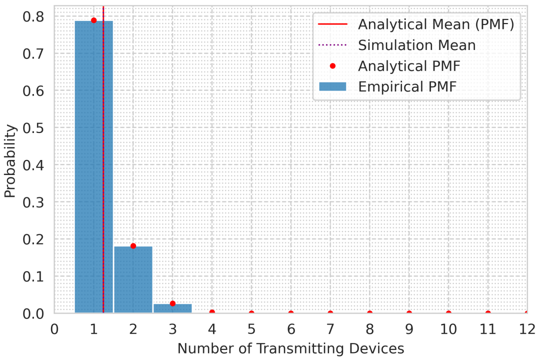
□



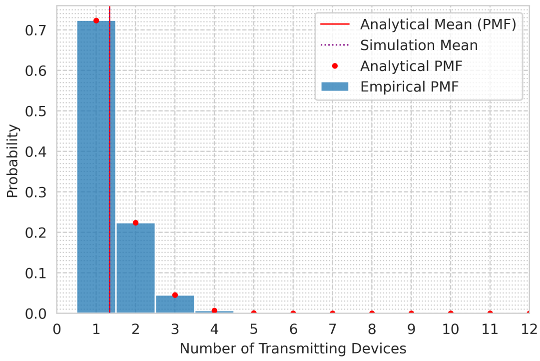
(a) K =10



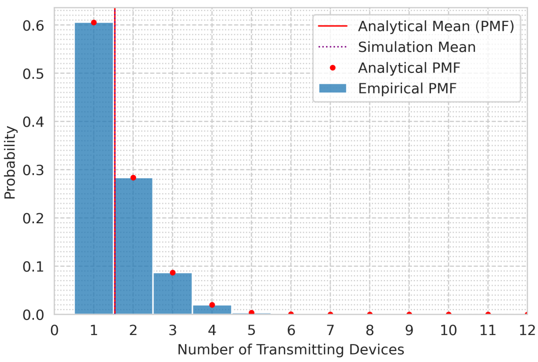
(b) $K = 20$



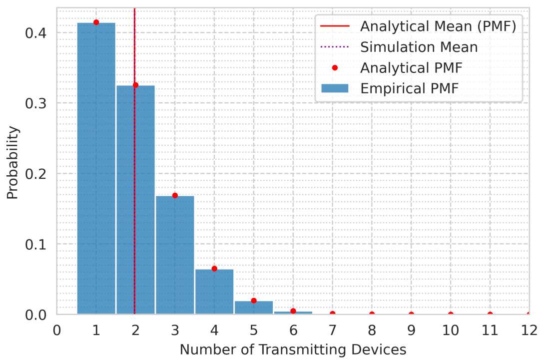
(c) $K = 30$



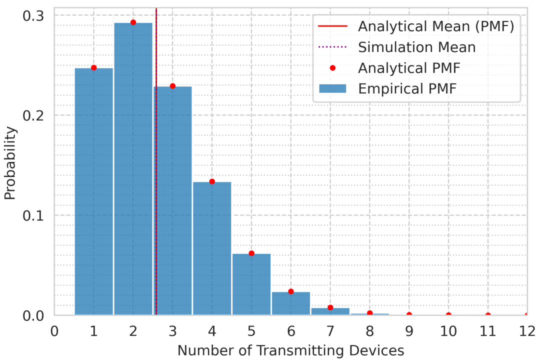
(d) $K = 40$



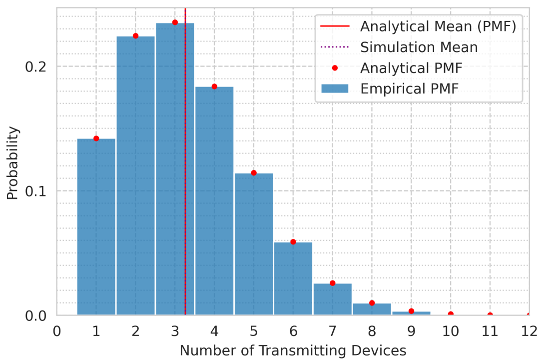
(e) $K = 60$



(f) $K = 100$



(g) $K = 150$



(h) $K = 200$

Figure 8: Probability mass function and expected value of the number of transmitting devices per detected preamble for $r = 64$.

To demonstrate the accuracy of the derived probability distribution and expected value, simulation and analytical results are presented. Simulation (empirical) results were obtained via the Monte Carlo method with 10^6 iterations, whereas the analytical ones were obtained by employing Theorem 6.1 expressions.

The distribution and expectation of the number of transmitting MTC devices per detected preamble $\left(W\right)$ is analyzed next. The mean value of W ranges from 1 with few MTC devices to around 3.5 with 200 MTC devices (Figure 8). The probability of two or more devices transmitting the same preamble (collision probability) increases as the number of devices increases. Note that for $K \leq 100$, a single MTC devices transmitting a detected preamble has the highest probability. For $K \leq 100$, the probability that two or fewer MTC devices are transmitting the same detected preamble is higher than 70%, whereas the probability that three or fewer MTC devices are transmitting the same detected preamble is higher than 90%. This low number of devices sharing the same preamble choice justifies employing techniques that depend on the number of devices transmitting the detected preambles, as the one being proposed in next section. Moreover, note that the analitycal results fit well the simulation ones.

6.2 The Probabilistic Retransmission Protocol

The proposed HARQ protocol is illustrated in Figure 9 and works as follows. Once an MTC device receives the RAR message containing an uplink grant to send the MSG3 message, the first MSG3 message is always transmitted. If this transmission is successfully decoded, the base station sends back an acknowledgement (ACK) message on the physical HARQ indicator channel (PHICH). The ACK in the first transmission indicates that just one MTC device was transmitting on the allocated resources (neglecting the capture effect) and the HARQ process is finalized as usual in both the UE and the base station. However, if the first transmission is not decoded, the base station sends back a negative-acknowledgment (NACK) message. This indicates that more than one MTC device transmitted on the allocated resources, or that a decoding error occurred due to channel impairments. Note that the no reception of an ACK message is also considered as a NACK message event.

As it is not possible to differentiate between the two above-mentioned cases, we propose to perform MSG3 retransmissions considering the two possibilities.

The proposed probability for uth MSG3 message transmission of a random access attempt initiated in ith RAO is denoted by $P_{\text{TX}iu}$ and defined as:

$$P_{\text{TX}iu} = \begin{cases} 1, & \text{if } u = 1 \\ \underbrace{P_{\text{RTX}} \cdot P_{iu}(\text{Collision}|\text{NACK})}_{\text{collision}} + \underbrace{1 \cdot P_{iu}(\text{Channel}|\text{NACK})}_{\text{channel impairements}}, & \text{if } u \geq 2 \end{cases}, \tag{15}$$

where

$$P_{iu}(\text{Channel}|\text{NACK}) = \frac{P_{iu}(\text{Channel} \cap \text{NACK})}{P_{iu}(\text{NACK})} = \frac{{p_e}^{(u-1)} \times P_{S,i}}{{p_e}^{(u-1)} \times P_{S,i} + P_{C,i}} \tag{16}$$

and

$$P_{iu}(\text{Collision}|\text{NACK}) = \frac{P_{iu}(\text{Collision} \cap \text{NACK})}{P_{iu}(\text{NACK})} = \frac{P_{C,i}}{{p_e}^{(u-1)} \times P_{S,i} + P_{C,i}} \tag{17}$$

are, respectively, the conditional probabilities that the retransmission comes from an uncollided preamble (Channel) or a collided preamble (Collision) transmission given that the MTC device received a NACK message[2] (NACK) in the $(u-1)$th transmission of a random access attempt initiated in the ith RAO. The probability of a noncollided preamble $P_{S,i}$ can be calculated as

$$P_{S,i} = \left(1 - \frac{1}{d_i}\right)^{K_i - d_i}, \tag{18}$$

where K_i and d_i are, respectively, the number of devices transmitting preambles and the number of detected preambles in the ith RACH occasion. The probability of a collided preamble is

$$P_{C,i} = 1 - P_{S,i} \tag{19}$$

[2]The NACK message is received on the PHICH for non-adaptive retransmissions or indicated through a DCI 0 message on the PDCCH for adaptive retransmissions.

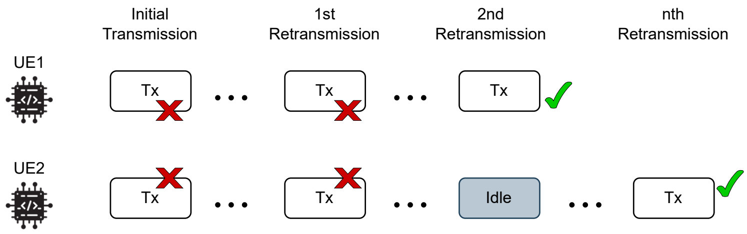


Figure 9: Illustration of the proposed probabilistic retransmission protocol. The UE_1 and UE_2 selected the same preamble at the first step of the random access procedure, thus they both receive the same random access response message. Their initial MSG3 transmission collides (a data collision event). In the first retransmission, both devices decide to retransmit, resulting in a data collision event again. In the second retransmission, however, UE_1 does transmit, whereas UE_2 does not, thus allowing the base station to successfully decode the transmission from UE_1. The base station can successfully decode the UE_2 transmission at the uth retransmission.

and the probability of error of a HARQ transmission p_e is a parameter of the radio access technology standard, usually defined as 0.1 [4]. Finally, P_{RTX} is the retransmission probability given that a collided preamble generated the NACK message received.

If the NACK message was caused by a channel error, an event which has a probability $P_{iu}(\text{Channel}|\text{NACK})$ to occur, the optimal retransmission probability equals 1. However, if the NACK message was caused by simultaneous transmission of various MTC devices on the same PUSCH resources, *i.e.*, a collision, with probability $P_{iu}(\text{Collision}|\text{NACK})$, the retransmission probability is P_{RTX}. Note that the optimal value for P_{RTX} is the inverse of the number of collided MTC devices within the same preamble. However, the MTC device does not know this value based only on the HARQ feedback. Moreover, this value is even unknown for the base station. Thus, in the next subsection, we introduce two policies to determine the P_{RTX} value by exploring the information available on the device side.

In the proposed HARQ protocol, if the first transmission of a MSG3 message is unsuccessful, the base station will reserve PUSCH resources for the remaining transmissions (retransmissions), regardless of the successful detection of a transmission for a given HARQ process. In this way, resources for performing the probabilistic retransmissions are guaranteed and more than one MTC may be detected with the same RAR message. This generates significant radio and energy resource savings since the number of RA attempts is reduced and the network access latency decreased. Note that this approach maintains the same resource utilization as the legacy HARQ process due to the collision of the MSG3 message re/transmission from devices that chose the same preamble. Such overhead will be analyzed numerically in Section 8.2.

Upon reception of an ACK message for its HARQ process, the MTC device determines if this ACK message is addressed to it by verifying the status of its last transmission. If it did not perform a transmission in the last scheduled HARQ opportunity (*i.e.,* if the generated random number was greater than P_{TX}), the ACK message is intended to another MTC devices that selected the same preamble. Thus, the MTC device continues with the HARQ process as described above. Otherwise, the MTC device finishes the HARQ process unilaterally.

6.3 Retransmission Policies

We propose two policies for defining the P_{RTX} value in (15) as follows.

6.3.1 Policy 1

The first policy (Policy 1) defines that

$$P_{\text{RTX}1} = \frac{1}{W_i + 1}, \tag{20}$$

where $W_i = K_i/d_i$ is the mean number of transmitting MTC devices per detected preamble in the ith RAO in which the MTC device performed the preamble transmission. This policy uses the fact that the MTC devices trying retransmission is a collided MTC device. Thus, $W_i + 1$ is used to obtain the retransmission probability in (20). Every time an ACK is received on the PHICH for the HARQ process of the MTC device in consideration, it indicates that an MSG3 message from other MTC devices in the same collision set was successfully decoded.

Thus, if the current W_i value is greater than one, we update the W_i value by subtracting the already decoded MTC device, $W_i = W_i - 1$.

6.3.2 Policy 2

The second policy (policy 2) is based on the number of MTC devices in collision with an MTC device that performs retransmission due to preamble collision. Let $Q_i = K_i - d_i$ be the maximum number of MTC devices that may be in a collision with a device that transmitted a detected preamble. Let O_i denote the number of MTC devices that choose the same preamble given that all the decoded preamble are already selected by exactly one MTC device in the ith RAO. $P_{\text{RTX}2}$ is defined as:

$$P_{\text{RTX}2} = \sum_{o=1}^{Q_i} \frac{1}{o+1} \cdot P(O=o)\,, \tag{21}$$

where $P(O = o)$ is the probability that o among Q_i MTC devices selected an already selected preamble, which follows a binomial distribution:

$$P(O=o) = \binom{Q_i}{o} \left(\frac{1}{d_i}\right)^{o} \left(1 - \frac{1}{d_i}\right)^{Q_i - o}\,. \tag{22}$$

To calculate $P_{\text{TX}iu}$ in (15), the variables unknown to the MTC device are K_i and d_i. Next, we provide a lightweight and standard-compatible method to estimate those variables in MTC devices.

6.4 Estimation of the Network Load

Existing methods in the literature for making this estimation are designed to operate at the base station side [15, 47–51]. Moreover, most of them involve recursive probability calculation and optimization problems. As the estimation of K_i, denoted by $\hat{K}_i$, needs to be performed by the MTC devices, we propose a simple but still efficient way based on the method proposed by Oh *et al.* [49], by adapting their proposal to the information available at the device side.

Let r_i be the number of available preambles for the contention-based random access procedure in the ith RAO. Let I_i and d_i be the number of unused preambles and decoded preambles as observed by the base station in the ith RAO, respectively. Even though several proposals assume that an base station is able to differentiate between a collided preamble and a non-collided preamble, this assumption does not hold in a real base station implementation [22]. Therefore,

$$r_i = I_i + d_i\,. \tag{23}$$

The probability of idle preambles in the ith RAO can be computed based on the observed I_i value as follows [49]:

$$\tilde{p}_{idle,i} = \frac{I_i}{r_i}. \tag{24}$$

This probability can also be calculated as [49]:

$$p_{idle,i} = \left(1 - \frac{1}{r_i}\right)^{K_i}, \tag{25}$$

and by using the first-order Taylor polynomial of $e^y \approx 1 + y$, which works well when y is small, $p_{idle,i}$ can be approximated as

$$p_{idle,i} \approx e^{-K_i/r_i}, \tag{26}$$

By setting $\tilde{p}_{idle,i}$ equals $p_{idle,i}$, the expected number of concurrent MTC devices in the ith RAO can be estimated as

$$\hat{K}_i = r_i \cdot \ln\left(\frac{r_i}{I_i}\right)\,. \tag{27}$$

By employing (23) in (27), $\hat{K}_i$ can be rewritten in terms of the d_i as

$$\hat{K}_i = r_i \cdot \ln\left(\frac{r_i}{r_i - d_i}\right)\,. \tag{28}$$

While the r_i value is always known by the devices through the system information broadcasted periodically, d_i is unknown to the MTC devices.

6.5 Estimation of the Number of Detected Preambles

The estimation method for the network load described in Section 6.4 was originally designed to operate in the base station, which has direct access to I_i and d_i variables at every transmission time interval (TTI). However, the devices do not have such information. To deal with this problem, we propose a technique to obtain d_i in the MTC device next.

6.5.1 The RAR Message Counting Technique

The proposed technique is based on the counting of the RAR messages sent by the base station after preambles detection. As all devices that performed preamble transmission in a given RACH occasion monitor the PDCCH for a possible match with its own RA-RNTI within the RAR window, we take advantage of this fact to perform a counting of the number of RAR messages presented in the PDSCH matching its RA-RNTI during the entire RAR window size. To guarantee that all RAR messages sent are included in the counting, the MTC devices keep performing the RAR counting until the end of the RAR window size, even if it receives a RAR message before the ending of the RAR window. Thus, the resulting counting is equivalent to the value of d_i. We call this, the RAR message counting rechnique.

6.5.2 Estimation under Inaccurate Information

The base station generally tries to respond to all decoded preambles by prioritizing RAR messages at the PDCCH scheduler and by allocating more PUSCH resources for sending RAR messages [52]. However, in some cases the base station may not be able to respond to all preamble decoded in a RAO because the resources available in a RAR window may not be sufficient to send RAR messages of all decoded preambles. For instance, the 3GPP proposed an MTC performance evaluation methodology in [4], which considers that for a 5 MHz cell with RAR window size of 5 ms, the maximum number of RAR messages per RAO equals 15. Thus, when the number of decoded preambles is higher than this value, the MTC devices are not able to make an accurate estimation of K_i from RAR message counting technique proposed above. This inaccurate information is more likely to happen in low bandwidth cells with high RACH loads.

For this reason, we propose a complementary method to obtain the estimated value, based on the access probability p_{ACB} of standardized ACB random access scheme, which was proposed for dealing with the radio access network (RAN) overload problem and signaling storms [4]. Generally, the ACB scheme is activated in a cell independent of the traffic types it supports. The base station periodically broadcasts the p_{ACB} value in the SIB2 [4]. With this probability, the ACB scheme limits the number of users trying random access under high signaling and RACH loads. The optimal p_{ACB} value which maximizes the RACH throughput was derived in [53] as:

$$p^{*}_{\text{ACB}} = \min\left(1, \frac{r}{K}\right), \tag{29}$$

where r is the number of available preambles for the contention-based random access procedure and K is the number of contending MTC devices, as estimated by the base station for the current ACB broadcasting period.

As both r and p_{ACB} are known from the SIB2, the MTC device is able to calculate the estimated value of K from the p_{ACB} received periodically by using (29) when p_{ACB} is less than 1. When $p_{\text{ACB}} = 1$, however, K can be estimated from (27). The calculation of $\hat{K}$ should be obtained from (29) when p_{ACB} is less than 1 because the base station has much hardware resources and more accurate information than the MTC devices. Note that in current cellular networks, the ACB scheme coexists with other specific solutions for MTC scenarios, such as the EAB scheme.

7 Performance Evaluation Methodology

We assess the performance of the proposed solutions via extensive simulations by using the LTE simulator (LTE-Sim) [31].

To simulate the system model in Section 3, we extended the LTE-Sim simulator [54]. We implemented the contention-based random access procedure [32], different random access schemes and traffic models proposed by the 3GPP in [4], the PDCCH resource management model proposed in Section 4, the two-stage approach for the packet scheduling [55], QoS support for uplink transmissions [55], and the padding and regular buffer status reports [32, 33]. We also implemented the non-contention based random access procedure [32, 33], and the support of random access traffic from handover users and HTC users. We also implemented a realistic PRACH power

Table 4: Simulation parameters for evaluation of the preamble-priority-aware scheme.

Parameter	Value
Number of contention-free preambles	12
Number of contention-based preambles	52
PRACH configuration index	6
Preamble format	0
RRS preamble sets	30 preambles for MTC devices 22 preambles for HTC users
Maximum number of preamble transmissions	10
Number of uplink grants per MAC RAR PDU	3 (PPA scheme) or 6 (HARQ Protocol)
ra-ResponseWindowSize	5 ms
mac-ContentionResolutionTimer	48 ms
Backoff indicator	20 ms
Maximum number of MSG3 transmissions	5
Error probability for lth preamble transmission attempt	e^{-l}
Error probability for MSG3 transmissions	0.1
PDCCH CCE resources	16
PDCCH aggregation level for MSG2	8 [2]
PDCCH aggregation level for MSG4	2
λ_{HO}	2.7 HO/s
λ_{HTC}	12 HTC/s
Maximum device transmit power	23 dBm
Received target power on PRACH	-118 dBm
Power ramping step	4 dBm

control mechanism, following the specifications in [32, 33]. The introduced random access implementations was extensively validated and the validation of part of this implementation can be found in [56].

We compare the random access performance of the preamble-priority-aware approach for scheduling control messages to that of the random access prioritized algorithm (Section 4.3) under resource constraints on the data and control channels. To assess the effect of PDCCH constraints with different random access schemes, we employed the traditional LTE random access scheme with both non-contention and contention-based operation modes as well as the RACH resource separation RAN overload control scheme, which was proposed to alleviate the negative effect of massive MTC on HTC. Since both preamble-priority-aware and random access prioritized scheduling algorithms use the principle of the MSG2-first policy, actual data transmission of user plane data is not necessary to assess their random access performance.

We compare the performance of the proposed probabilistic retransmission approach with p_{rx1} and p_{rx2} retransmission policies to that of the traditional random access procedure with the conventional HARQ protocol. The probabilistic retransmission protocol was evaluated employing the random access prioritized PDCCH algorithm. In addtition to the aforementioned implementations in the simulator, the access class barring RAN overload mechanism and the energy consumption capabilities were introduced in the simulator. The UE Lauridsen *et al.*'s energy consumption model [57] was employed.

7.1 Simulation Setup

A cellular network with support to HTC and MTC devices, as described in Section 3, is considered. The cell radius of the target cell was set to 0.5 km to simulate an urban macro-cell scenario. A base station with 5 MHz cell bandwidth in the FDD mode is located at the center of the cell with several devices uniformly distributed around it. Thus, 25 physical resource blocks are available in the cell [29] in each direction, from which one PRB is used for the PUCCH. There are $N_{\text{CCEs}}^{\text{PDCCH}} = 20$ CCEs available in the cell every subframe, and the fraction of control resources assigned to the PDCCH is assumed as 0.8 [30]. Each UE acts either as MTC device or as HTC user.

The number of HTC users attached to the cell was fixed to 300 user equipments. We assumed that around 40 percent of these users (120 UEs) have VoIP traffic, using the semi-persistent scheduling with initial random access. Therefore, $\lambda_{HTC} = \frac{120\,\text{HTC}}{10s} = 12\,\text{HTC/s}$. Based on the total number of RRC-connected devices, the mean number of X2-based (inter-BS) handovers per second is 1.8 [58]. To account for intra-BS handovers also, we add 50 percent to the calculated value of X2-based handovers. Therefore, $\lambda_{HO} = 2.7\,\mathrm{HO/s}$, which is a typical value in current cellular networks [59].

Scenarios with $5,000$ (*light load*), $10,000$ (*medium load*), and $30,000$ (*high load*) MTC devices were executed to assess the performance of the PPA scheme, as proposed by the 3GPP in [4]. In the traditional RA scheme, a set of 12 preambles is used by handover users for non-contention based random access and the remaining 52 preambles are shared between MTC devices and HTC users for CBRA. In the RRS scheme, the contention-based preambles are further separated between MTC devices (30 preambles) and HTC users (22 preambles).

In the evaluation of the probabilistic retransmission approach, six RAR messages are available per TTI and 52 preambles are shared among the MTC devices for executing the contention-based RA procedure and scenarios with $5,000$ (low load) and $10,000$ (medium load) MTC devices were executed, following the 3GPP MTC scenarios in [4][3].

Activation of MTC devices follows the $Beta(3,4)$ distribution within a 10 s interval to simulate an extreme scenario with MTC transmissions highly synchronized as also proposed by the 3GPP in [4].

We assume a MSG3 transmission error probability p_e of 0.1 [4]. Thus, any MSG3 message transmission from just one user can be successfully decoded with probability $1 - p_e$. Moreover, we assume that the transmission of MSG3 messages received from two or more users using the same radio resources cannot be decoded.

Table 4 summarizes the main configuration parameters used in these simulations.

7.2 Performance Metrics for the PPA Approach

The performance of the PPA proposal was assessed by using four metrics grouped into two perspectives: *network access performance* and *random access latency*. The *network access performance* perspective explores the capacity of the proposed solution to support random access differentiation by considering the access success and MSG2 blocking ratios, whereas the *random access latency* perspective evaluates the performance of the solution by highlighting its impact on the overall random access delay and MSG2 delays. These metrics are defined as follows.

7.2.1 Network access performance

- *Access Success Ratio*: the ratio of devices that successful perform the random access procedure.
- *MSG2 Blocking Ratio*: the number of MSG2 messages blocked divided by the total number of MSG2 messages that joined the MSG2 queue at the base station random access manager.

7.2.2 Random access latency

- *MSG2 Delay*: the average time between a preamble transmission and the reception of the corresponding RAR messages.
- *Random Access Delay*: the average time taken between the starting of the random access procedure and the successful reception of the message that finishes the corresponding random access procedure. Thus, it is averaged only among those devices that successfully finished the random access procedure. For the non-contention based random access mode, the reception of the RAR message finishes the random access procedure, whereas the reception of the device identity in a CR message finishes the random access procedure in the contention-based random access mode.

7.3 Performance Metrics for the Probabilistic Retransmission Protocol

The performance of the proposed probabilistic retransmission protocol was assessed by using the following seven metrics evaluating different perspectives of the proposed solution. They evaluate the RA performance, including access latency, RACH and overall random access procedure efficiency. Energy consumption and protocol overhead are also considered.

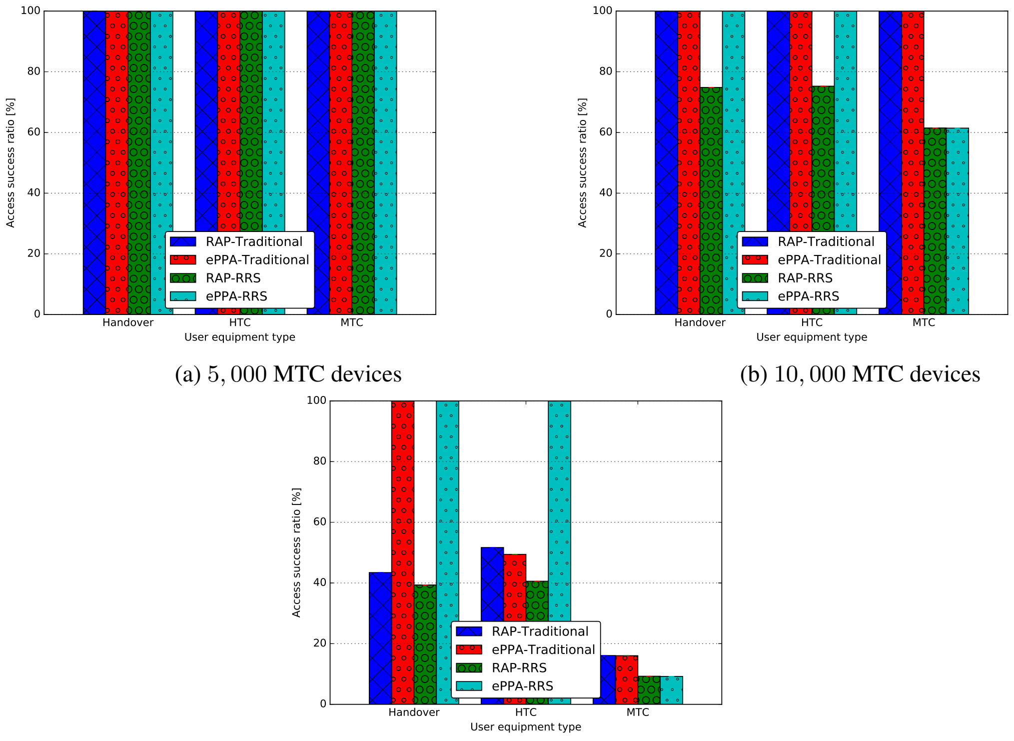
- *Average access delay*: the average time taken between the first preamble transmission and the successful reception of the device identity in a CR message finishes the corresponding random access procedure. Thus, it is averaged only among those devices that successfully finished the random access procedure.

[3]Since both random access and actual transmissions are involved in the evaluation of the probabilistic retransmission approach, the scenario with 30.000 was prohibitive to be executed due to computational capacity limitations.

- *Average number of successful MSG3 transmissions per RAR message*: the average number of devices successfully finishing the random access procedure with a RAR message.
- *RACH success probability*: the ratio between the number of successful devices in a RACH occasion (those receiving successfully the MSG4 message) and the total number of devices transmitting preambles in that RACH occasion.
- *Average number of collided MSG3 transmissions per device:* the total number of MSG3 transmitted divided by the total number of devices in the scenario.
- *Average number of transmitted preambles:* the average number of preambles transmitted to conclude the random access procedure. It is averaged among all the users in the scenario.
- *Average energy consumption per random access procedure:* the average amount of energy consumed by a device to perform the random access procedure. It is averaged among all the devices in the scenario.
- *HARQ protocol overhead on PUSCH*: the number of the number of PUSCH resources used in the HARQ process divided by the total number of devices in the scenario.

8 Simulation Results and Discussion

The figures in this section show mean values derived by using the independent replication method with 10 replication. Confidence intervals of 95 % confidence level are also shown. Each simulated scenario run until all started random access procedures finish, either successfully or unsuccessfully. The proposed preamble-priority-aware approach and probabilistic retransmission protocol are evaluated in Section 8.1 and Section 8.2, respectively.

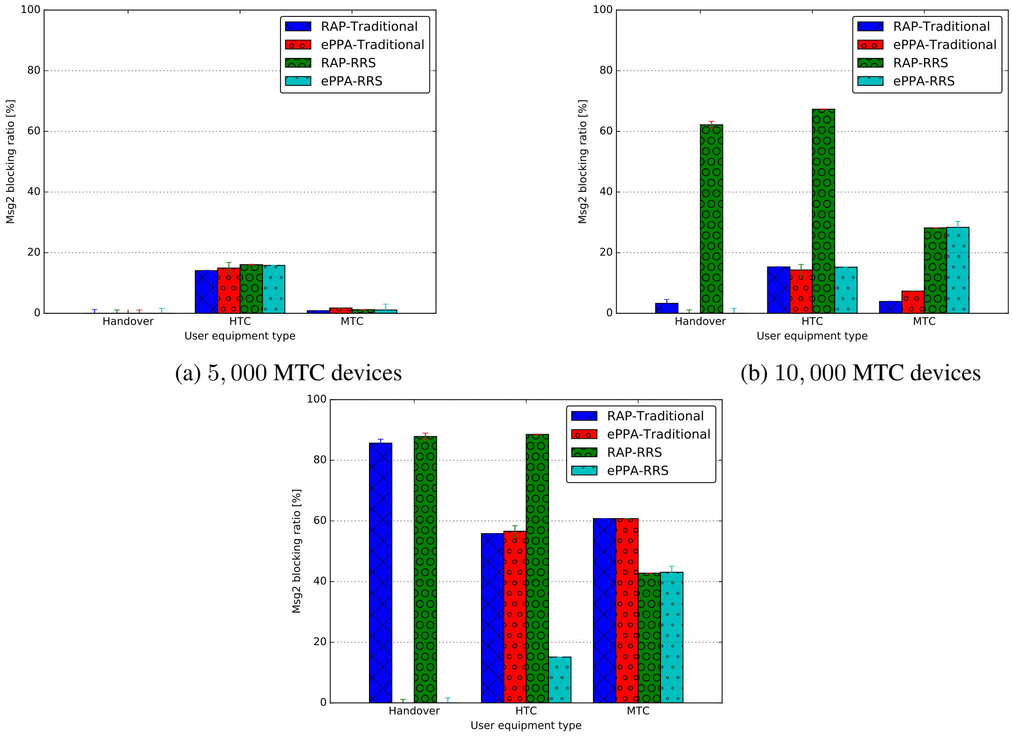


(a) 5, 000 MTC devices

(b) 10, 000 MTC devices

(c) 30, 000 MTC devices

Figure 10: Access success ratio for different number of MTC devices in the network. PPA: preamble-priority-aware; RAP: random access prioritized; RRS: RACH resource separation.



(a) $5,000$ MTC devices

(b) $10,000$ MTC devices

(c) $30,000$ MTC devices

Figure 11: MSG2 blocking ratio for different number of MTC devices in the network.

8.1 The Preamble-priority-aware Approach

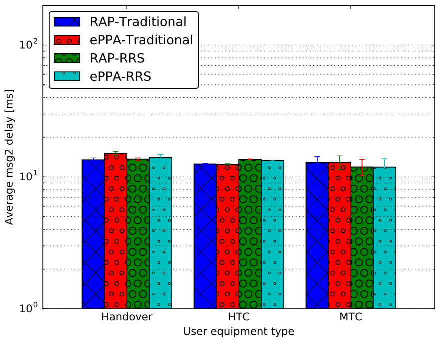
The access success ratio is shown in Figure 10. Under *light loads*, none of the evaluated scenarios has blocking of access requests (Figure 10a) since there are sufficient resources for the demand of all type of users (Figure 11a). Under *medium loads*, none of the traffic types presented blocking of access requests when the traditional RA scheme is employed (Figure 10b). However, under *high loads*, the PPA algorithm with the traditional RA scheme yields no loss of access attempts for handovers users but does not provide QoS guarantees for HTC users (Figure 10c). This is explained by the fact that handover users utilize the contention-free random access procedure, whereas HTC users share the contention-based preambles with MTC devices in the traditional RA scheme. Thus, the PPA algorithm gives high priority to control messages for handover users but does not prioritize control messages for HTC users over those for MTC devices.

On the other hand, a portion of the users utilizing prioritized preamble sequences is not able to get access to the network when the random access prioritized algorithm with RACH resource separation scheme is employed under higher loads, *i.e.*, $10,000$ (Figure 10b) and $30,000$ MTC devices (Figure 10c). This fact shows that the random access prioritized algorithm does not prioritize control messages derived from prioritized preambles.

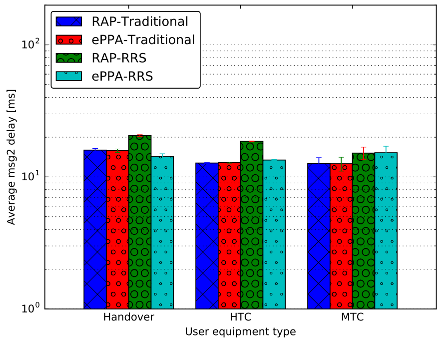
Due to the reduction of the number of preambles reserved for MTC devices in the RACH resource separation scheme, a high number of collisions of preamble transmissions from MTC device occurs. The MSG2-related resource utilization increases with the collision probability, since unsuccessful MSG3 transmissions implies on new preamble transmissions and more messages arriving at the MSG2 queue. As the resources for control messages is shared between MTC devices and HTC users and the random access prioritized algorithm does not differentiate MSG2 messages for any type of users, high MSG2 blocking ratio occurs for all type of users under *medium* (Figure 11b) and *high loads* (Figure 11c).

Conversely, the preamble-priority awareness approach of the proposed algorithm leads to no loss of access attempts for users with high and medium priority when the RACH resource separation scheme is used. The preamble-priority-aware algorithm is thus able to provide QoS differentiation during the random access procedure, even under massive access attempts.

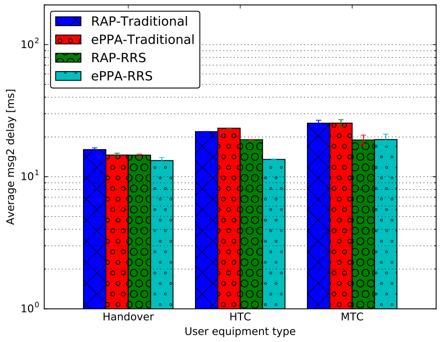
Figure 12 shows the average MSG2 delay. The average MSG2 delays for HTC users and MTC devices are quite similar. However, the PPA algorithm produces slightly lower MSG2 delays to user utilizing high priority preambles



(a) $5,000$ MTC devices



(b) $10,000$ MTC devices



(c) $30,000$ MTC devices

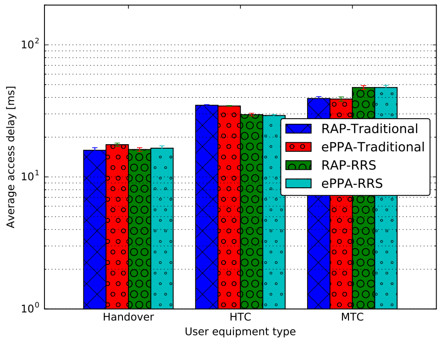
Figure 12: Average MSG2 delay for different number of MTC devices in the network.

sequences (handover users) than does the random access prioritized (RAP) algorithm with both conventional RA scheme and RRS scheme. Moreover, this is achieved with a low increment in the MSG2 message delays for the other users. This is explained by the prioritization applied by the PPA approach to the non-contention based random access preambles, which receive the highest priority.

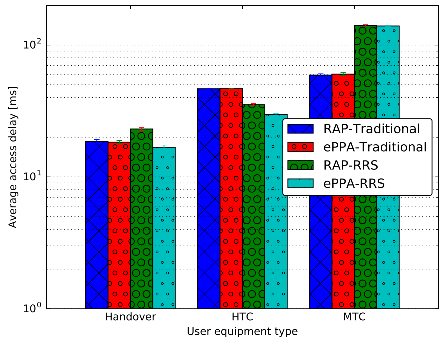
Figure 13 shows the average access delay. As expected, the access delays of handover users are shorter than those of the VoIP users and MTC devices. This is mainly due to the fact that handover users utilize contention-free random access procedure, whereas the latter two users utilize contention-based random access procedure. For the UEs using contention-based random access procedure with $5,000$ MTC devices, access delays obtained by employing both random access schemes are quite similar, whereas access delays given by the RACH resource separation scheme for $10,000$ MTC devices are larger than those given by the conventional random access scheme. This is because preambles sequences reserved for MTC devices are reduced, increasing the collision probability of these UEs. For handover users, these values are quite similar to that of the MSG2 delay because of the contention-free random access procedure used by these users. For users utilizing contention-based preambles sequences, the transmission of the MSG3 message may collide when two or more users transmit the same preamble sequence, thus increasing the overall access delay. This collision is usually detected only when the MSG4 message is not received by the UE. As the average access delay is calculated only for UEs which random access procedures are successfully finished, it cannot produce further gain due to the use of a preamble-priority awareness approach in the PDCCH scheduling algorithm when high random access channel collision and shortages of PDCCH resources sporadically occur.

8.2 The Probabilistic Retransmission Protocol for the Random Access Procedure

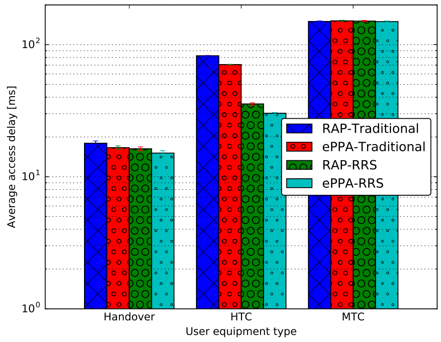
The proposed approach achieves lower average access delays than does the conventional RA scheme in all settings (Figure 14). Under low loads, the proposed approach gives access delays between $2\,\%$ (Policy 2) and $10\,\%$ (Policy 1) lower than the conventional RA scheme. Under medium loads, the access delay improvement provided by our proposal is even more significant, yielding access delays between $20\,\%$ (Policy 2) and $30\,\%$ (Policy 1) lower than those of the conventional scheme. This significant decrease in the access delay occurs because the proposed probabilistic HARQ retransmission strategy considerably reduces the number of collided MSG3 message



(a) 5, 000 MTC devices



(b) 10, 000 MTC devices



(c) 30, 000 MTC devices

Figure 13: Average random access delay for different number of MTC devices in the network.

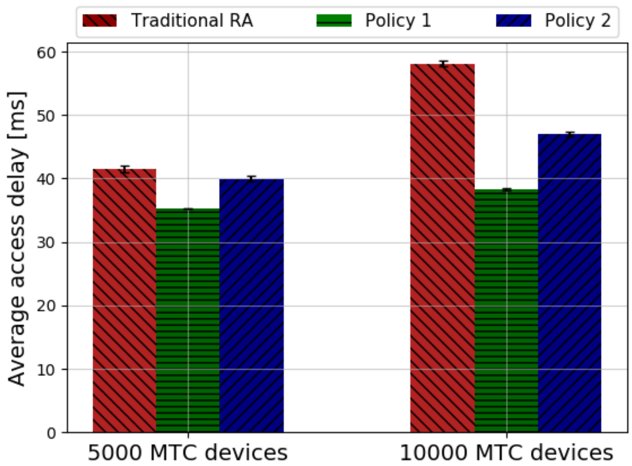


Figure 14: Average access delay

transmissions (Figure 15) and increases the number of successful MSG3 transmissions per issued RAR message (Figure 16).

One of the most important results obtained with our proposal is the reduction in the number of collided MSG3 transmissions (Figure 15). Under all loads, the proposed approach reduces in more than 75 % the collided MSG3 transmissions when compared to the conventional RA scheme. This is a direct result of applying the probabilistic retransmission strategy to the MSG3 transmissions in the random access procedure. Note that, even though the retransmission Policy 2 yields a slightly lower number of collided MSG3 transmissions than does the Policy 1, the Policy 2 produces higher access delays and lower number of successful MSG3 transmissions per issued RAR message. This is explained by the fact that the Policy 1 updates the expected number of collided MTC devices per detected preamble with every received ACK message, whereas the Policy 2 maintains the same value during all retransmissions. This continuous updating adjusts the retransmission probability based on the HARQ feedback for a given process, making the Policy 1 less aggressive than Policy 2, which is calculated just once each RA attempt. This also evinces that the updating of the retransmission policy impacts positively on the performance of the probabilistic retransmission approach. Moreover, the decreasing in the collided MSG3 transmissions induced by the proposed approach yields lower average number of transmitted preambles per successful random access procedure when compared to the traditional RA scheme (Figure 17). It also increases the RACH success probability for a given RACH occasion (Figure 18). In fact, RACH success probability values yield by Policy 1 increases as the number of MTC devices also increases, thus indicating that our technique increases its efficiency as the RACH load increases. Conversely, the efficiency of the traditional random access procedure decreases as the number of MTC devices increases.

Under low loads, the Policy 1 gives a number of successful MSG3 transmissions per RAR message 10 % higher than that given by the conventional RA scheme and the Policy 2, whereas under medium loads, the proposed policies outperform the conventional RA scheme, providing a number of successful MSG3 transmissions per RAR message 10 % (Policy 2) and 25 % (Policy 1) higher than the conventional RA scheme. This is also a result of the probabilistic strategy of our proposal. The adjustment of the retransmission probability performed by the Policy 1 also explains this difference between the two policies.

Moreover, the decrease in the average access delay (Figure 14), collided MSG3 transmissions (Figure 15), and number of transmitted preambles (Figure 17) produced by the proposed approach yields lower device energy consumption for the whole random access procedure when compared to the conventional scheme (Figure 19). Under low loads, our proposal gives from 5% (Policy 2) to 10% (Policy 1) energy saving when compared to energy consumed by the conventional RA scheme. Under medium loads, the energy savings increases between 20% (Policy 2) and 25% (Policy 1). These results show the incapacity of the conventional RA scheme to deal with the *MSG3 collision problem*, which significantly impacts on the energy consumption of the MTC devices. Moreover, as this is an important key performance indicator for the MTC devices given their limited battery capacity and prohibitive costs incurred with battery replacement in most cases, our approach is relevant for future IoT scenarios over cellular networks. Besides, the above-discussed gains yielded by our proposal are obtained with lower PUSCH resource overhead than that of the conventional RA scheme (Figure 20). Under low loads, the PUSCH resources utilized by our proposal with retransmission Policy 1 is about 5 % lower than that utilized by the conventional scheme, whereas the Policy 2 gives almost the same performance as the conventional scheme. However, under medium loads, the proposed scheme achieves between 10% (Policy 2) to 25% (Policy 1) lower PUSCH resource utilization when compared to those utilized by the conventional scheme. Again, the retransmission Policy 1 yields less HARQ overhead because of its update every time an ACK message is received from the base station for the specific HARQ process, which reduces the number of RA attempts.

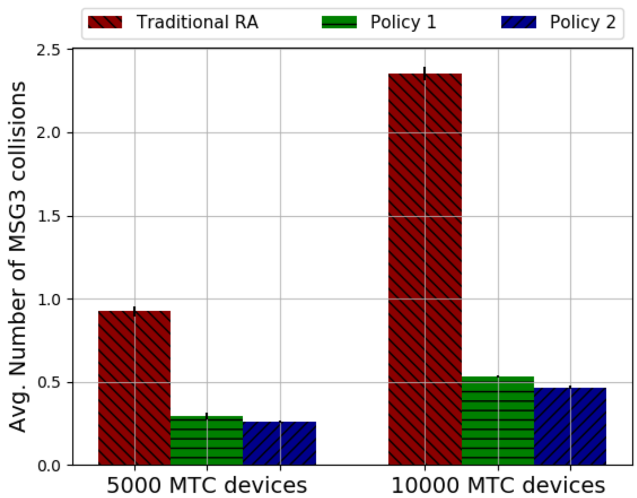


Figure 15: Average number of collided MSG3 transmissions per device

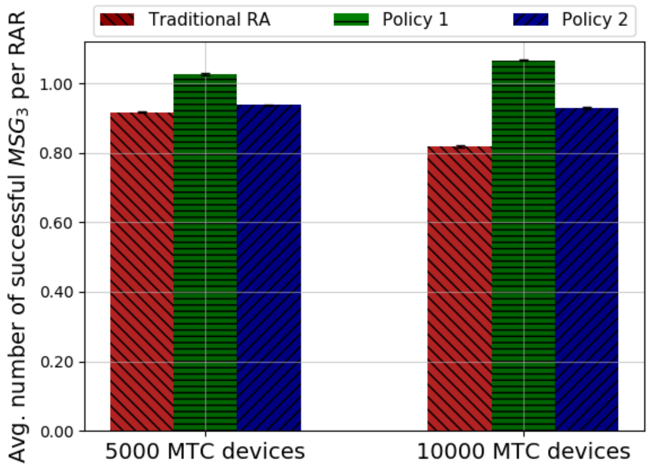


Figure 16: Average number of successful MSG3 transmissions per RAR message.

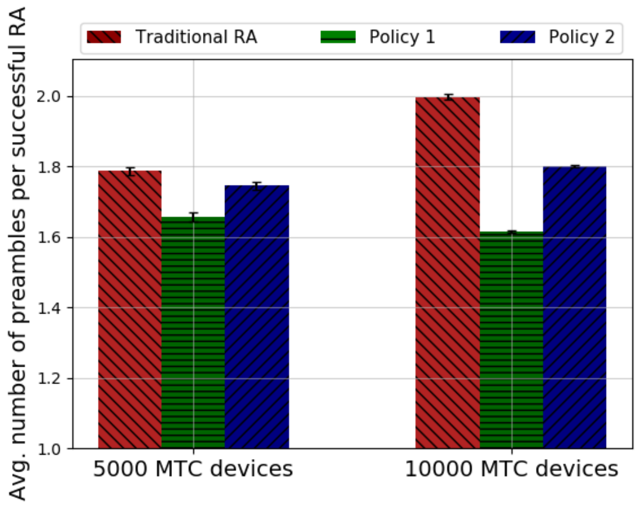


Figure 17: Average number of transmitted preambles

9 Conclusion

This article proposes and evaluates medium access control solutions for the support of massive machine-type communications in 3GPP cellular networks by exploring the inter playing between random access and resource allocation. Particularly, the *shortage of random access resources* and the problems derived from it were addressed. First, a resource allocation mechanisms to deal with the effects of the *control message blocking* on network performance is presented. Second, a probabilistic retransmission approach to ameliorate the *problem of data collision* in the random access procedure of cellular Internet of things network technologies is introduced.

The former proposed a novel scheme for scheduling random access control messages in cellular networks supporting massive machine-type communications. The proposal introduces the concept of preamble-priority awareness by taking into consideration the priority of the preamble that derived the control messages to be scheduled. In the preamble-priority-aware approach, the scheduling prioritization is based on the explicit or implicit preamble priorities defined by the underline random access scheme employed in the network by the mobile network operator. Control scheduling with preamble-priority awareness is particularly relevant for scenarios in which radio resources are shared among various device or traffic types, and shortage of control and data channel

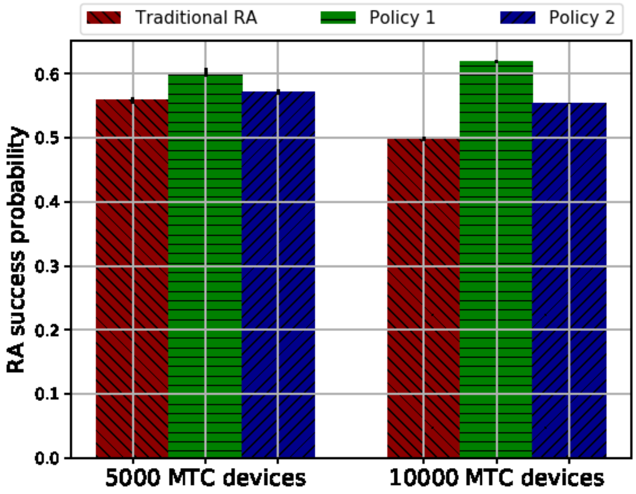


Figure 18: RACH success probability.

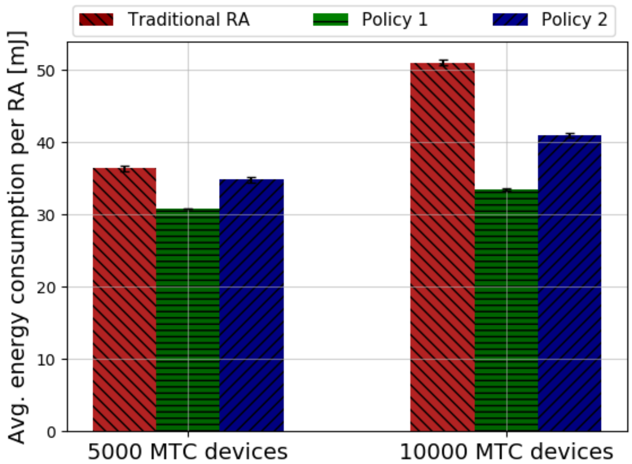


Figure 19: Average energy consumption per random access procedure

resources may occur because of a massive number of devices accessing the network simultaneously. A first example of such scenarios is a HTC/MTC coexistence supported by the same network. It is also relevant for devices requiring low-latency and network access guarantees, such as in the case of handover or applications executing real-time applications. This concept could also be explored for RAN virtualization, in which resource isolation is one of the key requirements [60].

The latter proposed a probabilistic retransmission approach to ameliorate the *problem of data collision* in the random access procedure of cellular Internet of things network technologies. Specifically, we introduced the application of a probability value for every MSG3 retransmission to increase the chance of MSG3 successful decoding under preamble collisions. Two different retransmission policies were proposed to decrease the MSG3 collisions; one based on the mean number of collided MTC devices per detected preamble and another based on the number of MTC devices in collision with an MTC device that performs retransmission due to a preamble collision. To apply the proposed approach at the MTC device side, a novel method to estimate the number of MTC devices trying random access in a given random access opportunity is proposed. This method is based on random access response message counting and the probability broadcasted by the base station in the standardized access class barring random access scheme.

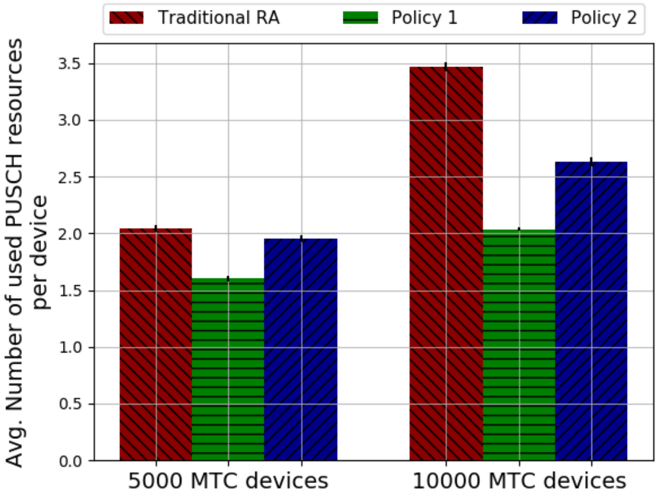


Figure 20: HARQ protocol overhead on the PUSCH.

Simulation results showed that the preamble-priority-aware approach, under radio resource limitations due to heavy random access channel load, is able to increase the chances of accessing the network and reduce the access delay for devices whose preambles are prioritized by the underline random access scheme in the initial phase of the random access procedure. Moreover, even though the RACH resource separation scheme has been considered a good option to isolate the impact of MTC on HTC [4], existing literature on RAN overload control schemes has neglected the resource constraints imposed by control and data channels on the performance. The results presented here, which take into account these constraints, evince that the RACH resource separation scheme only achieves its intended objective with the help of the preamble-priority-aware approach. On the other hand, results on the proposed probabilistic retransmission protocol showed that the it effectively reduces the number of MSG3 collisions. Moreover, our proposal saves energy, reduces the access delay, and yet reduces the channel utilization when compared to the conventional random access scheme.

Thus, the proposed solutions are fundamental to improve the support of massive machine-type communications in cellular Internet of things network technologies.

References

[1] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, “Internet of things: A survey on enabling technologies, protocols, and applications,” *IEEE Communications Surveys Tutorials*, vol. 17, no. 4, pp. 2347–2376, Fourthquarter 2015.

[2] M.-Y. Cheng, G.-Y. Lin, H.-Y. Wei, and A.-C. Hsu, “Overload control for machine-type-communications in LTE-advanced system,” *Communications Magazine, IEEE*, vol. 50, no. 6, pp. 38–45, June 2012.

[3] A. Laya, C. Kalalas, F. Vazquez-Gallego, L. Alonso, and J. Alonso-Zarate, “Goodbye, ALOHA!” *IEEE Access*, vol. 4, pp. 2029–2044, 2016.

[4] 3GPP, “Technical Specification Group Radio Access Network; Study on RAN Improvements for Machine-type Communications,” 3rd Generation Partnership Project (3GPP), TR 37.868, September 2011.

[5] GSMA, “Mobile IoT in the 5G future - NB-IoT and LTE-M in the context of 5G,” White Paper, May 2018.

[6] M. Series, “Detailed specifications of the terrestrial radio interfaces of international mobile telecommunications-2020 (IMT-2020),” *Recommendation ITU*, vol. 2150, p. 255, 2021.

[7] H. Jin, W. T. Toor, B. C. Jung, and J.-B. Seo, “Recursive pseudo-bayesian access class barring for M2M communications in LTE systems,” *IEEE Transactions on Vehicular Technology*, vol. 66, no. 9, pp. 8595–8599, 2017.

[8] P. Bertrand and J. Jiang, “Random access,” in *LTE – The UMTS Long Term Evolution*, 2nd ed., S. Sesia, I. Toufik, and M. Baker, Eds. John Wiley & Sons, Ltd, 2011, pp. 371–406.

[9] C. A. Astudillo, T. P. C. de Andrade, and N. L. S. da Fonseca, “Allocation of control resources with preamble

priority awareness for human and machine type communications in LTE-Advanced networks," in *2017 IEEE International Conference on Communications (ICC)*, Paris, France, May 2017, pp. 1–6.

[10] ——, "Impact of preamble-priority-aware downlink control signaling scheduling on LTE/LTE-A network performance," in *IEEE 86th Vehicular Technology Conference (VTC2017-Fall)*, Toronto, Canada, Sep. 2017, pp. 1–5.

[11] T. P. C. de Andrade, C. A. Astudillo, and N. L. S. da Fonseca, "Impact of M2M traffic on human-type communication users on the LTE uplink channel," in *7th IEEE Latin-American Conf. Commun. (LATINCOM)*, November 2015, pp. 1–6.

[12] ——, "Allocation of control resources for machine-to-machine and human-to-human communications over lte/lte-a networks," *IEEE Internet of Things Journal*, vol. 3, no. 3, pp. 366–377, June 2016.

[13] C. A. Astudillo, F. H. S. Pereira, and N. L. S. da Fonseca, "Probabilistic retransmissions for the random access procedure in cellular IoT networks," in *2019 IEEE International Conference on Communications*, May 2019, pp. 1–7.

[14] M. Condoluci, M. Dohler, G. Araniti, A. Molinaro, and J. Sachs, "Enhanced radio access and data transmission procedures facilitating industry-compliant machine-type communications over LTE-based 5G networks," *IEEE Wireless Commun.*, vol. 23, no. 1, pp. 56–63, Feb. 2016.

[15] S. Vural, N. Wang, P. Bucknell, G. Foster, R. Tafazolli, and J. Muller, "Dynamic preamble subset allocation for ran slicing in 5g networks," *IEEE Access*, vol. 6, pp. 13 015–13 032, 2018.

[16] T. Kim, K. S. Ko, and D. K. Sung, "Prioritized random access for machine-to-machine communications in OFDMA based systems," in *2015 IEEE Int. Conf. Commun. (ICC)*, June 2015, pp. 2967–2972.

[17] ——, "Prioritized random access for accommodating M2M and H2H communications in cellular networks," in *2015 IEEE Global Commun. Conf. (GLOBECOM) Workshops*, December 2015, pp. 1–6.

[18] M. Condoluci, M. Dohler, G. Araniti, A. Molinaro, and J. Sachs, "Enhanced radio access and data transmission procedures facilitating industry-compliant machine-type communications over LTE-based 5G networks," *IEEE Wireless Commun.*, vol. 23, no. 1, pp. 56–63, Feb. 2016.

[19] M. S. Ali, E. Hossain, and D. I. Kim, "LTE/LTE-A random access for massive machine-type communications in smart cities," *IEEE Communications Magazine*, vol. 55, no. 1, pp. 76–83, January 2017.

[20] M. Vilgelm, S. R. Linares, and W. Kellerer, "Enhancing cellular M2M random access with binary countdown contention resolution," in *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Oct 2017, pp. 1–6.

[21] M. Vilgelm and W. Kellerer, "Binary contention resolution for M2M random access prioritization in LTE-A and 5G," in *2017 IFIP Networking Conference (IFIP Networking) and Workshops*, June 2017, pp. 1–2.

[22] J. S. Kim, D. Munir, S. F. Hasan, and M. Y. Chung, "Enhancement of LTE RACH through extended random access process," *Electronics Letters*, vol. 50, no. 19, pp. 1399–1400, September 2014.

[23] D. Magrin, C. Pielli, C. Stefanovic, and M. Zorzi, "Enabling LTE RACH collision multiplicity detection via machine learning," *CoRR*, vol. abs/1805.11482, 2018. [Online]. Available: http://arxiv.org/abs/1805.11482

[24] Z. Alavikia and A. Ghasemi, "A multiple power level random access method for M2M communications in LTE-A network," *Transactions on Emerging Telecommunications Technologies*, vol. 28, no. 6, p. e3137, 2017.

[25] J. Kim and J. Lee, "Exploiting the capture effect to enhance RACH performance in cellular-based M2M communications," *Sensors*, vol. 17, no. 10, 2017.

[26] K. S. Ko, M. J. Kim, K. Y. Bae, D. K. Sung, J. H. Kim, and J. Y. Ahn, "A novel random access for fixed-location machine-to-machine communications in ofdma based systems," *IEEE Communications Letters*, vol. 16, no. 9, pp. 1428–1431, September 2012.

[27] Y. Liang, X. Li, J. Zhang, and Z. Ding, "Non-orthogonal random access for 5G networks," *IEEE Trans. on Wireless Commun.*, vol. 16, no. 7, pp. 4817–4831, 2017.

[28] P. V. Orlik and S. S. Rappaport, "On the handoff arrival process in cellular communications," *Wirel. Networks*, vol. 7, no. 2, pp. 147–157, 2001. [Online]. Available: https://doi.org/10.1023/A:1016685506058

[29] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical channels and modulation," 3rd Generation Partnership Project (3GPP), TS 36.211, October 2013.

[30] F. Capozzi, D. Laselva, F. Frederiksen, J. Wigard, I. Kovacs, and P. Mogensen, "UTRAN LTE downlink system performance under realistic control channel constraints," in *Vehicular Technology Conference Fall (VTC 2009-Fall), 2009 IEEE 70th*, Sept 2009, pp. 1–5.

[31] G. Piro, L. A. Grieco, G. Boggia, F. Capozzi, and P. Camarda, "Simulating LTE cellular systems: An open-source framework," *IEEE Trans. Veh. Technol.*, vol. 60, no. 2, pp. 498–513, 2011.

[32] 3GPP, "Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) protocol

specification," 3rd Generation Partnership Project (3GPP), TS 36.321, July 2013.
[33] ——, "5G; NR; Medium Access Control (MAC) protocol specification," 3rd Generation Partnership Project (3GPP), TS 38.321, January 2022.
[34] ——, "Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Rresource Control (RRC); Protocol specification," 3rd Generation Partnership Project (3GPP), TS 36.331, September 2013.
[35] ——, "5G; NR; Radio Rresource Control (RRC) Protocol specification," TS 38.331, January 2022.
[36] ——, "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN) overall description stage 2," 3rd Generation Partnership Project (3GPP), TS 36.300, January 2022.
[37] D. Hong and S. S. Rappaport, "Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures," *IEEE Trans. Veh. Technol.*, vol. 35, no. 3, pp. 77–92, Aug. 1986.
[38] D. Jiang, H. Wang, E. Malkamaki, and E. Tuomaala, "Principle and performance of semi-persistent scheduling for VoIP in LTE system," in *Wireless Communications, Networking and Mobile Computing, 2007. WiCom 2007. International Conference on*, Sept 2007, pp. 2861–2864.
[39] S.-M. Oh, S. Cho, J.-H. Kim, and J. Kwun, "VoIP scheduling algorithm for AMR speech codec in IEEE 802.16e/m system," *IEEE Communications Letters*, vol. 12, no. 5, pp. 374–376, 2008.
[40] H. Lee, T. Kwon, and D.-H. Cho, "An enhanced uplink scheduling algorithm based on voice activity for VoIP services in IEEE 802.16d/e system," *IEEE Communications Letters*, vol. 9, no. 8, pp. 691–693, 2005.
[41] J. B. Seo and V. C. M. Leung, "Performance modeling and stability of semi-persistent scheduling with initial random access in LTE," *IEEE Trans. Wireless Commun.*, vol. 11, no. 12, pp. 4446–4456, 2012.
[42] F. Capozzi, D. Laselva, F. Frederiksen, J. Wigard, I. Kovacs, and P. Mogensen, "UTRAN LTE downlink system performance under realistic control channel constraints," in *Vehicular Technology Conference Fall (VTC 2009-Fall), 2009 IEEE 70th*, Sept 2009, pp. 1–5.
[43] O. Grøndalen, A. Zanella, K. Mahmood, M. Carpin, J. Rasool, and O. N. Østerbø, "Scheduling policies in time and frequency domains for LTE downlink channel: A performance comparison," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 4, pp. 3345–3360, 2017.
[44] P. Osti, P. Lassila, S. Aalto, A. Larmo, and T. Tirronen, "Analysis of PDCCH performance for M2M traffic in LTE," *Vehicular Technology, IEEE Transactions on*, vol. 63, no. 9, pp. 4357–4371, November 2014.
[45] S. Sesia, I. Toufik, and M. Baker, Eds., *Bertrand, Pierre and Jiang, Jing*, 2nd ed. John Wiley & Sons, Ltd, 2011.
[46] S. Tekinay and B. Jabbari, "Handover and channel assignment in mobile cellular networks," *IEEE Communications Magazine*, vol. 29, no. 11, pp. 42–46, 1991.
[47] Y. C. Pang, S. L. Chao, G. Y. Lin, and H. Y. Wei, "Network access for M2M/H2H hybrid systems: a game theoretic approach," *IEEE Communications Letters*, vol. 18, no. 5, pp. 845–848, 2014.
[48] G. Lin, S. Chang, and H. Wei, "Estimation and adaptation for bursty LTE random access," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 4, pp. 2560–2577, April 2016.
[49] C. Oh, D. Hwang, and T. Lee, "Joint access control and resource allocation for concurrent and massive access of M2M devices," *IEEE Transactions on Wireless Communications*, vol. 14, no. 8, pp. 4182–4192, Aug 2015.
[50] M. Tavana, V. Shah-Mansouri, and V. W. S. Wong, "Congestion control for bursty M2M traffic in LTE networks," in *2015 IEEE International Conference on Communications (ICC)*, June 2015, pp. 5815–5820.
[51] G. C. Madueño, N. K. P. andČ. Stefanović, and P. Popovski, "Massive M2M access with reliability guarantees in LTE systems," in *2015 IEEE International Conference on Communications (ICC)*, June 2015, pp. 2997–3002.
[52] W. Kim and E. Lee, "LTE network enhancement for vehicular safety communication," *Mobile Information Systems*, vol. 2017, pp. 8 923 782:1–8 923 782:18, 2017. [Online]. Available: https://doi.org/10.1155/2017/8923782
[53] S. Duan, V. Shah-Mansouri, Z. Wang, and V. W. S. Wong, "D-ACB: Adaptive congestion control algorithm for bursty M2M traffic in LTE networks," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 12, pp. 9847–9861, Dec 2016.
[54] G. Piro, "LTE-Sim - the LTE simulator," http://telematics.poliba.it/LTE-Sim, 2011 (accessed April 17, 2015).
[55] C. A. Astudillo, J. Freitag Borin, and N. L. S. da Fonseca, "LTE time-domain uplink scheduler for QoS provisioning," in *Global Communications Conference (GLOBECOM), 2013 IEEE*, Dec 2013, pp. 1458–1463.
[56] T. P. C. de Andrade, C. A. Astudillo, L. R. Sekijima, and N. L. S. da Fonseca, "The random access procedure in long term evolution networks for the Internet of Things," *IEEE Communications Magazine*, vol. 55, no. 3,

March 2017.
[57] M. Lauridsen, L. Noël, T. B. Sørensen, and P. Mogensen, "An empirical LTE smartphone power model with a view to energy efficiency evolution," *Intel Technology Journal*, vol. 18, no. 1, pp. 172–193, 2014.
[58] R. Lilius, J. Salo, J. M. T. Pérez, and E. M. Metsälä, *Planning and Optimizing Mobile Backhaul for LTE*. John Wiley and Sons, Ltd, 2015, pp. 129–237.
[59] A. Elnashar and M. El-saidny, *Practical Guide to LTE-A, VoLTE and IoT - Paving the Way Towards 5G*. John Wiley & Sons, Ltd, 2018.
[60] R. Kokku, R. Mahindra, H. Zhang, and S. Rangarajan, "NVS: A substrate for virtualizing wireless resources in cellular networks," *IEEE/ACM Transactions on Networking*, vol. 20, no. 5, pp. 1333–1346, Oct 2012.