Universitat Politècnica de València

Departamento de Comunicaciones

Ph.D. Thesis

«On reliable and energy efficient massive wireless communications: the road to 5G»

Author: Israel Leyva-Mayorga

Advisors: Dr. Vicent Pla
Dr. Jorge Martinez-Bauset

Valencia
December 2018

To my wife, Mónica,
and to my parents,
because they have
given me the most
precious gifts I can
think of: their love,
confidence, and
support.

Acknowledgment

First of all I would like to thank my wife, Mónica, for all her support and confidence throughout this journey. I am extremely grateful she is part of my life as I could not ask for a better companion during my adventures. A great deal of the success I have been able to achieve is due to her.

I would also like to thank my parents, who have given me their support and confidence since my early stages. I am completely sure this is the reason I am who I am right now. My success is just a reflection of yours as parents.

Naturally, I am greatly thankful to my supervisors, Vicent and Jorge, for everything they have taught me, directly and indirectly, during these years. My adventure in the Ph.D. started with a visit to this same lab four years ago and my admiration and respect for them has been increasing ever since. For instance, I am still impressed on the time and dedication they devote to us when writing papers and obtaining results, but also with administrative musts. I will be eternally thankful.

Likewise, I also thank the support and hospitality provided by Prof. Dr.-Ing. Dr. h. c. Frank H. P. Fitzek and Dr. Rico Radeke before, during, and after my research stay at the Technische Universität Dresden (TU Dresden). In this regard, I would also like to thank all the people involved in making this stay possible. These include people from the Deutscher Akademischer Austauschdienst (DAAD) and the ERASMUS+ offices at the UPV and TU Dresden, whose monetary support was essential.

I would also like to express my gratitude to my main collaborators, Luis Tello Oquendo at the UPV and Roberto Torre at the TU Dresden. Hopefully we will soon meet again.

Lastly, I would like to thank the support provided by the Consejo Nacional de Ciencia y Tecnología, México (CONACYT) and by the Consejo Mexiquense de Ciencia y Tecnología (COMECYT) through grant CONACYT-GEM 383936 for postgraduate studies in a foreign country. This includes the support for my short stays at the Universidad Autónoma del Estado de México (UAEM). These stays were only possible due to the support provided by Prof. Otniel Portillo Rodríguez. Thanks for the invitation.

Abstract

The 5th generation (5G) of mobile networks is just around the corner. It is expected to bring extraordinary benefits to the population and to solve the majority of the problems of current 4th generation (4G) systems. The success of 5G, whose first phase of standardization has concluded, relies in three pillars that correspond to its main use cases: massive machine-type communication (mMTC), enhanced mobile broadband (eMBB), and ultra-reliable low latency communication (URLLC). This thesis mainly focuses on the first pillar of 5G: mMTC, but also provides a solution for the eMBB in massive content delivery scenarios. Specifically, its main contributions are in the areas of: 1) efficient support of mMTC in cellular networks; 2) random access (RA) event-reporting in wireless sensor networks (WSNs); and 3) cooperative massive content delivery in cellular networks.

Regarding mMTC in cellular networks, this thesis provides a thorough performance analysis of the RA procedure (RAP), used by the mobile devices to switch from idle to connected mode. These analyses were first conducted by simulation and then by an analytical model; both of these were developed with this specific purpose and include one of the most promising access control schemes: the access class barring (ACB). To the best of our knowledge, this is one of the most accurate analytical models reported in the literature and the only one that incorporates the ACB scheme. Our results clearly show that the highly-synchronized accesses that occur in mMTC applications can lead to severe congestion. On the other hand, it is also clear that congestion can be prevented with an adequate configuration of the ACB scheme. However, the configuration parameters of the ACB scheme must be continuously adapted to the intensity of access attempts if an optimal performance is to be obtained. We developed

a practical solution to this problem in the form of a scheme to automatically configure the ACB; we call it access class barring configuration (ACBC) scheme. The results show that our ACBC scheme leads to a near-optimal performance regardless of the intensity of access attempts. Furthermore, it can be directly implemented in 3rd Generation Partnership Project (3GPP) cellular systems to efficiently handle mMTC because it has been designed to comply with the 3GPP standards. This combination of characteristics is rarely present in other access control schemes reported in the literature.

In addition to the analyses and the solution described above for cellular networks, a general analysis for smart metering applications is performed. That is, we study an mMTC scenario from the perspective of event detection and reporting WSNs. Specifically, we provide a hybrid model for the performance analysis and optimization of cluster-based RA WSN protocols. Results obtained with this model showcase the utility of overhearing to minimize the number of packet transmissions, but also of the adaptation of transmission parameters after a collision occurs. Building on this, we are able to provide some guidelines that can drastically increase the performance of a wide range of RA protocols and systems in event reporting applications.

Regarding eMBB, we focus on a massive content delivery scenario in which the exact same content is transmitted to a large number of mobile users simultaneously. Such a scenario may arise, for example, with video streaming services that offer a particularly popular content. This is a problematic scenario because cellular base stations have no efficient multicast or broadcast mechanisms. Hence, the traditional solution is to replicate the content for each requesting user, which is highly inefficient. To solve this problem, we propose the use of network coding (NC) schemes in combination with cooperative architectures named mobile clouds (MCs). Specifically, we develop a protocol for efficient massive content delivery, along with the analytical model for its optimization. Results show the proposed model is simple and accurate, and the protocol can lead to energy savings of up to 37 percent when compared to the traditional approach. In addition, the proposed solution sharply reduces the cellular data consumed by the mobile devices.

Resumen

La quinta generación de redes móviles (5G) se encuentra a la vuelta de la esquina. Se espera que esta nueva generación provea de beneficios extraordinarios a la población y, también, que resuelva la mayoría de los problemas de las redes de cuarta generación (4G) actuales. El éxito de 5G, cuya primera fase de estandarización ha sido completada, depende de tres pilares; cada uno de ellos corresponde a uno de sus casos de uso: comunicaciones tipo-máquina masivas, banda ancha móvil mejorada y comunicaciones ultra fiables y de baja latencia (mMTC, eMBB y URLLC, respectivamente, por sus siglas en inglés). En esta tesis nos enfocamos en el primer pilar de 5G, mMTC, pero también proveemos una solución para lograr eMBB en escenarios de distribución masiva de contenidos. Específicamente, las principales contribuciones son en las áreas de: 1) soporte eficiente de mMTC en redes celulares; 2) acceso aleatorio para el reporte de eventos en redes inalámbricas de sensores (WSNs); y 3) cooperación para la distribución masiva de contenidos en redes celulares.

En el apartado de mMTC en redes celulares, esta tesis provee un análisis profundo del desempeño del procedimiento de acceso aleatorio, que es la forma mediante la cual los dispositivos móviles acceden a la red. Estos análisis fueron inicialmente llevados a cabo por medio de simulaciones y, posteriormente, por medio de un modelo analítico. En ambos tipos de análisis los modelos fueron desarrollados específicamente para este propósito e incluyen uno de los esquemas de control de acceso más prometedores: *access class barring* (ACB). Nuestro modelo es uno de los más precisos que se pueden encontrar en la literatura y el único que incorpora el esquema de ACB. Los resultados obtenidos por medio de este modelo y por simulación son claros: los accesos altamente sincronizados que ocurren en aplicaciones de mMTC pueden causar congestión severa

en el canal de acceso. Por otro lado, también son claros en que esta congestión se puede prevenir con una adecuada configuración del ACB. Sin embargo, los parámetros de configuración del ACB deben ser continuamente adaptados a la intensidad de accesos para poder obtener un desempeño óptimo. En la tesis se propone una solución práctica a este problema en la forma de un esquema de configuración automática para el ACB; lo llamamos ACBC. Los resultados muestran que nuestro esquema puede lograr un desempeño muy cercano al óptimo sin importar la intensidad de los accesos. Asimismo, puede ser directamente implementado en redes celulares para soportar el tráfico mMTC, ya que ha sido diseñado teniendo en cuenta los estándares del *3rd Generation Partnership Project* (3GPP). Esta combinación de características difícilmente se encuentra en otros esquemas de control de acceso reportados en la literatura.

Además de los análisis descritos anteriormente para redes celulares, se realiza un análisis general para aplicaciones de contadores inteligentes. Es decir, estudiamos un escenario de mMTC desde la perspectiva de las WSNs con tareas de detección y reporte de eventos. Específicamente, desarrollamos un modelo híbrido para el análisis de desempeño y la optimización de protocolos de WSNs de acceso aleatorio y basados en *cluster*. Los resultados obtenidos por medio de este modelo muestran la utilidad de escuchar el medio inalámbrico para minimizar el número de transmisiones y también de modificar las probabilidades de transmisión después de una colisión. Con base en los resultados, somos capaces de proponer directrices que pueden mejorar drásticamente el desempeño de una amplia gama de protocolos y sistemas de acceso aleatorio para aplicaciones de reporte de eventos.

En lo que respecta a eMBB, nos enfocamos en un escenario de distribución masiva de contenidos, en el que un mismo contenido es enviado de forma simultánea a un gran número de usuarios móviles. Un escenario de este tipo puede ocurrir, por ejemplo, con servicios de *streaming* de vídeo que ofrecen un contenido particularmente popular. Este escenario es problemático, ya que las estaciones base de la red celular no cuentan con mecanismos eficientes de *multicast* o *broadcast*. Por lo tanto, la solución que se adopta comúnmente es la de replicar e contenido para cada uno de los usuarios que lo soliciten; está claro que esto es altamente ineficiente. Para resolver este problema, proponemos el uso de esquemas de *network coding* y de arquitecturas cooperativas

llamadas nubes móviles. En concreto, desarrollamos un protocolo para realizar la distribución masiva de contenidos de forma eficiente, junto con un modelo analítico para su optimización. Los resultados demuestran que el modelo propuesto es simple y preciso, y que el protocolo puede reducir el consumo energético hasta en un 37% con respecto al enfoque tradicional. Además, la solución propuesta reduce drásticamente los datos móviles consumidos por los dispositivos.

Resum

La cinquena generació de xarxes mòbils (5G) es troba molt a la vora. S'espera que aquesta nova generació proveïsca de beneficis extraordinaris a la població i, també, que resolga la majoria dels problemes de les xarxes de quarta generació (4G) actuals. L'èxit de 5G, per a la qual ja ha sigut completada la primera fase del qual d'estandardització, depén de tres pilars; cadascun d'ells correspon a un dels seus casos d'ús: comunicacions tipus-màquina massives, banda ampla mòbil millorada, i comunicacions ultra fiables i de baixa latència (mMTC, eMBB i URLLC, respectivament, per les seues sigles en anglés). En aquesta tesi ens enfoquem en el primer pilar de 5G, mMTC, però també proveïm una solució per a aconseguir eMBB en escenaris de distribució massiva de continguts. Específicament, les principals contribucions són en les àrees de: 1) suport eficient de mMTC en xarxes cel·lulars; 2) accés aleatori per al report d'esdeveniments en xarxes sense fils de sensors (WSNs); i 3) cooperació per a la distribució massiva de continguts en xarxes cel·lulars.

En l'apartat de mMTC en xarxes cel·lulars, aquesta tesi realitza una anàlisi profunda de l'acompliment del procediment d'accés aleatori, que és la forma mitjançant la qual els dispositius mòbils accedeixen a la xarxa. Aquestes anàlisis van ser inicialment dutes a terme per mitjà de simulacions i, posteriorment, per mitjà d'un model analític. En tots dos tipus d'anàlisi els models van ser desenvolupats específicament per a aquest propòsit i inclouen un dels esquemes de control d'accés més prometedors: el *access class barring* (ACB). El nostre model és un dels més precisos que es poden trobar en la literatura i l'únic que incorpora l'esquema d'ACB. Els resultats obtinguts per mitjà d'aquest model i per simulació són clars: els accessos altament sincronitzats que ocorren en aplicacions de mMTC poden causar congestió severa en el canal d'accés.

D'altra banda, també són clars en què aquesta congestió es pot previndre amb una adequada configuració de l'ACB. No obstant això, els paràmetres de configuració de l'ACB han de ser contínuament adaptats a la intensitat d'accessos per a poder obtindre unes prestacions òptimes. En la tesi es proposa una solució pràctica a aquest problema en la forma d'un esquema de configuració automàtica per a l'ACB; l'anomenem ACBC. Els resultats mostren que el nostre esquema pot aconseguir un acompliment molt proper a l'òptim sense importar la intensitat dels accessos. Així mateix, pot ser directament implementat en xarxes cel·lulars per a suportar el trànsit mMTC, ja que ha sigut dissenyat tenint en compte els estàndards del *3rd Generation Partnership Project* (3GPP). Aquesta combinació de característiques difícilment es troba en altres esquemes de control d'accés reportats en la literatura.

A més de les anàlisis descrites anteriorment per a xarxes cel·lulars, es realitza una anàlisi general per a aplicacions de comptadors intel·ligents. És a dir, estudiem un escenari de mMTC des de la perspectiva de les WSNs amb tasques de detecció i report d'esdeveniments. Específicament, desenvolupem un model híbrid per a l'anàlisi de prestacions i l'optimització de protocols de WSNs d'accés aleatori i basats en clúster. Els resultats obtinguts per mitjà d'aquest model mostren la utilitat d'escoltar el mitjà sense fil per a minimitzar el nombre de transmissions i també de modificar les probabilitats de transmissió després d'una col·lisió. Amb base en els resultats, som capaços de proposar directrius que poden millorar dràsticament l'acompliment d'una àmplia gamma de protocols i sistemes d'accés aleatori per a aplicacions de report d'esdeveniments.

Pel que fa a eMBB, ens enfoquem en un escenari de distribució massiva de continguts, en el qual un mateix contingut és enviat de forma simultània a un gran nombre d'usuaris mòbils. Un escenari d'aquest tipus pot ocórrer, per exemple, amb serveis de *streaming* de vídeo que ofereixen un contingut particularment popular. Aquest escenari és problemàtic, ja que les estacions base de la xarxa cel·lular no compten amb mecanismes eficients de *multicast* o *broadcast*. Per tant, la solució que s'adopta comunament és la de replicar el contingut per a cadascun dels usuaris que ho sol·liciten; és clar que això és altament ineficient. Per a resoldre aquest problema, proposem l'ús d'esquemes de *network coding* i d'arquitectures cooperatives anomenades núvols mòbils. En concret, desenvolupem un protocol per a realitzar la distribució massiva de continguts

de forma eficient, juntament amb un model analític per a la seua optimització. Els resultats demostren que el model proposat és simple i precís, i el protocol pot reduir el consum energètic fins a un 37% respecte a l'enfocament tradicional. A més, la solució proposada redueix dràsticament les dades mòbils consumides pels dispositius.

Contents

List of Acronyms

LTE-A LTE Advanced

2G 2nd generation

3G 3rd generation

3GPP 3rd Generation Partnership Project

4G 4th generation

5G 5th generation

5GPPP 5G Infrastructure Public Private Partnership

AB adaptive backoff

ACB access class barring

ACBC access class barring configuration

ALE adaptive line enhancer

ARQ automatic repeat request

BCM backoff CM

CCDF complementary CDF

CCE control channel element

CDF cumulative distribution function

CE coverage enhancement

CH cluster head

CM cluster member

DOF degree of freedom

DTMC discrete-time Markov chain

eMBB enhanced mobile broadband

eMBMS evolved multimedia broadcast multicast service

eNB evolved NodeB

ERM extended RM

FB fixed backoff

FDM frequency-division multiplexing

H2H human-to-human

HARQ hybrid ARQ

IFI idealized full state information

IoT Internet of Things

ITU International Telecommunication Union

JSD Jensen-Shannon Divergence

KPI key performance indicator

LMS least-mean-square

LTE Long Term Evolution

M2M machine-to-machine

MAC medium access control

MC mobile cloud

MIB Master Information Block

mMTC massive machine-type communication

MSE mean squared error

MTC machine-type communications

NB-IoT narrowband Internet of Things

NC network coding

NCC network-coded cooperation

OFDM orthogonal frequency-division multiplexing

OFDMA orthogonal frequency-division multiple access

PALE "pulling" ALE

PDCCH physical downlink control channel

pdf probability density function

PER packet erasure ratio

PH phase-type

pmf probability mass function

PRACH physical RACH

QoS quality of service

RA random access

RACH random access channel

RAN radio access network

RAO random access opportunity

RAP RA procedure

RAR RA response

RLNC random linear NC

RLS recursive least-squares

RM reference model

RTT round-trip time

RV random variable

SIB System Information Block

TDM time-division multiplexing

TDMA time-division multiple access

TM traffic model

UE user equipment

URLLC ultra-reliable low latency communication

WSN wireless sensor network

List of Figures

List of Tables

Chapter 1

Introduction

Society is on the edge of a technological revolution in which every daily-life object will soon be provisioned with processing and communication capabilities; the so-called Internet of Things (IoT). As such, novel applications are pushing the boundaries on the capacity of the current 4th generation (4G) of mobile networks. As a consequence, there has been an amazingly rapid increase in the number of wireless devices in recent years and it has been accompanied with a similarly rapid increase in wireless data traffic [32]. For instance, the projected number of wireless devices by 2020 is around 11.6 billion and the expected data traffic by 2021 is around 49 exabytes per month. Current cellular networks cannot handle such a large number of connections nor such a high amount of data traffic.

The 3rd Generation Partnership Project (3GPP) has recently concluded the first phase of standardization for the 5th generation (5G) of mobile networks. 5G promises improved data rates (in the order of a few Gbps), a greater capacity, lower latency, and, in general, a much better quality of service (QoS) when compared to 4G [14]. Therefore, 5G is expected to solve most, if not all, of the problems of 4G. In addition, 5G will provide a high level of integration by combining the new interface with 4G and with short-range technologies such as WiFi [19]. The combination of all the previously described characteristics envisions to provide users with ubiquitous and real-time access to information and services.

Since the introduction of digital wireless communications to the phone industry in 1991 with the 2nd generation (2G), subsequent generations have merely represented an incremental advance in technology [84]. That is, the 3rd generation (3G) and 4G simply increased the achievable data rate of the users, but neglected many more aspects of wireless communications. Instead, 5G is focused towards three main use cases, as defined by the International Telecommunication Union (ITU) and the 3GPP. These are the pillars of 5G: massive machine-type communication (mMTC), enhanced mobile broadband (eMBB), and ultra-reliable low latency communication (URLLC) [7, 49]. In this thesis, we focus on providing efficient solutions for the former use case: mMTC, but also investigate a technique to enhance performance in a specific eMBB scenario in Chapter 6.

The term mMTC stands for the autonomous exchange of data between a massive number of wireless devices. This novel paradigm enables a myriad of applications such as smart metering, fleet management, traffic optimization, e-health care, and vehicle control [6, 103]. As such, achieving efficient mMTC is essential to attain a resilient IoT. Needless to say, 4G technology was developed to handle human-to-human (H2H) traffic, the same as previous mobile generations, and has several limitations that make it inefficient to handle mMTC [58, 90].

One of the first studies that revealed some of the problems that may arise under mMTC applications in the current 4G LTE Advanced (LTE-A) networks was conducted by the 3GPP itself [1]. Specifically, the 3GPP considered an urban scenario in central London. In this urban scenario, each household within the coverage area of a macro-base station is equipped with one smart metering device, which is set to transmit its information directly to the cellular base station. Interestingly, the total number of households served by the cellular base station was found to be greater than 30 000. Therefore, it is no surprise that the study revealed severe congestion can occur if such a high number of devices attempt to access the cellular base station in a highly synchronized manner. This is a typical behavior in mMTC applications [1, 76, 103]. Hence, from that point on, exhaustive efforts have been made to support mMTC in LTE-A [37].

Despite its limitations, the current 4G LTE-A system will serve as a base to 5G.

For instance, the RA procedure (RAP) that is used by the mobile users to switch from idle to connected mode in LTE-A has been incorporated to the narrowband Internet of Things (NB-IoT) standard with only minor modifications. This standard was published by the 3GPP in release 13 and is a low-power wide-area implementation at the 4G cellular base stations that is considered a 5G technology [106]. The main purpose of NB-IoT is to support mMTC by providing a higher power efficiency than in typical LTE-A. Specifically, NB-IoT nodes are expected to be battery powered for as long as 10 years, to be connected to cellular base stations as far as 10 km, and to drastically reduce manufacturing costs [7]. Furthermore, the 5G Infrastructure Public Private Partnership (5GPPP) METIS-II project was in charge of designing a new radio access network (RAN) for 5G. In this new RAN, the exact same RAP is considered with one minor difference that allows high-priority devices to have a higher access probability [11]. On the other hand, the proposed solution to handle mMTC is to incorporate short-range technologies to create groups of mobile devices that contain a leader. Then, the leader performs the RAP in representation of the whole group. This fulfills the promise of the integration of the 5G interface with 4G and short-range technologies, but also confirms the prevalence of the RAP defined for LTE-A, at least for the coming years.

One of the main problems of the RAP is that its first step, preamble transmission, resembles a simple multichannel slotted ALOHA access protocol. For instance, orthogonal sequences are used in traditional LTE-A whereas orthogonal frequency patterns are used in NB-IoT. Then, a collision occurs if multiple devices transmit the same preamble simultaneously. The literature on slotted ALOHA protocols is vast and there is a clear consensus that this type of protocols are prone to congestion when the system capacity is exceeded. Cellular networks under mMTC applications, where highly synchronized accesses occur frequently, are no exception to this rule. Besides, the first step of the RAP is not its only limitation. Cellular base stations also present limitations on the downlink control channel, used for the second step of the RAP. This second message signals the success of an access attempt at the first step. As a consequence, an access attempt may fail due to: 1) a preamble collision; 2) insufficient downlink control resources; or 3) wireless channel errors.

Access control mechanisms are the most promising approach to efficiently support

mMTC in 4G and beyond. Among these mechanisms, the access class barring (ACB) is especially interesting as has already been included in the LTE-A standards [10] and provides a probabilistic mechanism to delay the access requests of mobile devices to reduce the signaling traffic intensity. The main focus of this thesis is on the performance analysis of the RAP, the evaluation of the potential benefits of the ACB scheme, and on the development of an adaptive scheme to correctly configure the ACB parameters.

The main contributions and scientific publications derived from this thesis with respect to mMTC in 3GPP cellular networks are described in the following. It presents one of the most thorough performance analyses of the RAP and of the ACB scheme that can be found in the literature. This analysis is first performed by simulation, where the problems of the RAP and benefits of the ACB scheme are exhibited [68, 95]. Next, it presents an analytical model of the RAP that includes the ACB scheme. To the best of our knowledge, this is one of the most thorough and accurate analytical models of the RAP that can be found in the literature and is the only one that incorporates the ACB scheme [60, 69]. Furthermore, this analytical model can be used to accurately assess the performance of the RAP under a wide range of network configurations and can be modified to incorporate different assumptions commonly made in the literature. Finally, this thesis presents an adaptive solution to congestion in the form of an access class barring configuration (ACBC) scheme. That is, it presents a mechanism to automatically configure the parameters of the ACB scheme according to the signaling traffic intensity. This mechanism is particularly valuable as it strictly adheres to the 3GPP standards and can yield a near-optimal performance regardless of the signaling traffic intensity [66, 67].

Besides the work oriented to support mMTC in cellular networks under the 3GPP standard, this thesis presents an analysis of mMTC from a more general perspective: wireless sensor networks (WSNs). A WSN is an auto-organized collection of nodes with wireless communication and environmental sensing capabilities. These nodes are in charge of collecting and transmitting information regarding the state of a particular physical parameter of interest [16]. Therefore, WSNs are cost-efficient solutions to massive monitoring and that are not, in general, subject to the 3GPP or to any particular standard. Specifically, we focus on random access (RA) event reporting in time-critical applications. These applications usually involve the detection of hazardous conditions,

so they have stringent time and reliability requirements [85].

Our work on WSNs focuses on the performance analysis and optimization of a RA protocol that incorporates two novel approaches. The first one is to set a the number of data packets required to characterize the occurring phenomena. Then, nodes overhear the wireless medium to identify the exact point in time when these packets are transmitted. At this point in time, nodes can drop packets that are still pending for transmission, which eliminates the transmission of redundant packets. The second approach is to modify transmission probabilities after a collision occurs (i.e., during backoff). While this approach is not entirely new, its combination with the first approach offers great benefits to the network. For instance, one of the main benefits of the combination of these two approaches is a dramatic increase in the robustness of the performance of the network to the inaccurate selection of parameters. That is, the network is capable of providing a near-optimal performance even when the selection of parameters is far from optimal. As a summary, the main contributions and scientific publications derived from this thesis with respect to WSNs include the formulation of the hybrid model, the proposal of a RA protocol for event reporting in critical-time applications that incorporates the approaches described above, and the performance evaluation and parameter optimization of this protocol [62, 63].

As described above, a specific scenario for the second of the three main use cases for 5G: eMBB, is also studied in this thesis. The term eMBB stands for the demand of high data rates across a wide coverage area, and is clearly associated to multimedia consumption in mobile devices. Some of the deployment scenarios for eMBB are indoor hot spots, high speed vehicles, virtual and augmented reality, and gaming. A scenario that is specially problematic is that of massive content delivery, in which the exact same content is transmitted to a large number of mobile users simultaneously. A clear example of such scenario arises with streaming services that offer a particularly popular content. In such case, the cellular base stations must either replicate the content for each mobile user or utilize inefficient broadcast implementations such as the evolved multimedia broadcast multicast service (eMBMS) [104]. Needless to say, content replication leads to an irrational waste of wireless resources whereas the eMBMS usually suffers from unexpected disconnections and lacks support to ensure an adequate QoS to individual mobile devices [26]. An especially promising solution

to this problem is to offload the cellular link by shifting the traffic to short-range links. The strength of such an approach greatly increases with the coming of 5G, where short and long-range technologies will complement each other.

The novel paradigm of network-coded cooperation (NCC) refers to the combination of network coding (NC) schemes with cooperative architectures known as mobile clouds (MCs). The benefits of NCC in the scenario described above were investigated during a research stay at the Deutsche Telekom Chair of Communication Networks of the Technische Universität Dresden, in Dresden, Germany. The research on NCC during this period comprised the formulation of a protocol for massive content delivery and the analytical model to optimize this protocol. Results show that this NCC protocol can provide significant energy and cellular data savings to the mobile devices, but also can lead to considerable throughput gains when compared to the traditional approach of replicating the content for each requesting user. As such, the main contributions of this thesis on eMBB are the formulation of the NCC protocol and of the model for its optimization [61].

The rest of this thesis is organized in six chapters. Chapters 2, 3, and 4 are dedicated to the support mMTC in cellular networks. Specifically, Chapters 2 and 3 present a thorough performance analysis of the RAP, including the ACB scheme as defined in the 3GPP technical specifications [5, 8, 10]. Chapter 2 presents results obtained by simulation. On the other hand, Chapter 3 presents an analytical model for the RAP that includes the ACB scheme; results obtained by this model are also presented in this chapter. Chapter 4 presents the ACBC scheme, which is the proposed solution to efficiently support mMTC in cellular networks. Next, Chapter 5 presents a hybrid method for the performance analysis of RA protocols in WSNs for event-reporting applications. Hence, it provides an approach to support mMTC that is independent from the 3GPP standards. Chapter 6 presents a novel NCC protocol for eMBB applications in cellular networks under a massive content delivery scenario. The analytical model that describes the operation of the NCC protocol is also presented and used to optimize its performance. Finally, Chapter 7 presents the main conclusions and promising lines of research.

Chapter 2

Performance analysis of RA in cellular networks under mMTC scenarios

2.1 Introduction

The novel communication paradigm of massive machine-type communication (mMTC) is one of the use major use cases for the 5th generation (5G) of mobile networks and stands for the autonomous exchange of data between an exceedingly large number of wireless devices; these mobile devices are known as machine-to-machine (M2M) user equipments (UEs). MMTC enables a wide range of applications that are appealing to both the academy and industry such as smart metering, fleet management, e-health, and many more. Due to the proliferation of mMTC applications, the number of deployed M2M UEs is growing at an incredibly rapid pace [32].

The current 4th generation (4G) LTE Advanced (LTE-A) system has a widely deployed infrastructure, which provides with ubiquitous coverage and global connectivity [2, 72]. As such, LTE-A networks present nowadays one of the best solutions for the interconnection of mobile devices and will serve as a foundation for the development of the 5G system [24, 29, 76]. In fact, 5G networks are expected to provide efficient support for mMTC.

Nevertheless, cellular technology up to 4G was developed to handle human-to-human (H2H) traffic, where few UEs (compared to the billions of M2M devices expected by 2020 [32]) communicate simultaneously and transmit relatively large amounts of data. Conversely, in mMTC, a bulk of M2M UEs communicate sparingly with cellular base stations, known as evolved NodeBs (eNBs) in 4G, in a highly synchronized manner [76, 103]. While the data packets sent in machine-type communications (MTC) are small in size when compared to the size of data packets in H2H communications, the large number of access requests may exceed the signaling capacity of the eNBs. This phenomenon leads to severe network congestion and to the loss of potentially critical information [37, 58].

The UEs access the eNB by means of the RA procedure (RAP); it is performed through the random access channel (RACH) and comprises a four-message handshake: preamble transmission (only allowed in predefined time/frequency resources called random access opportunities (RAOs), RA response (RAR), connection request, and contention resolution messages. The RAP defined by the 3rd Generation Partnership Project (3GPP) for both, 4G and 5G, is described in detail in Section 2.3. Still it is important to point out that the main bottlenecks of the RAP are in the first two messages: preamble transmission and RAR. The reason for this is that uplink resources for preamble transmission and downlink resources for the RAR are limited and shared by every UE within the cell. On the other hand, connection request and contention resolution messages are sent through dedicated resources.

Specifically, the number of preambles are selected randomly by the accessing UEs. Hence, collisions occur when multiple UEs select and transmit the same preamble at the same RAO. On the other hand, RAR messages are transmitted by the eNB and contain a limited number of *uplink grants*, each of which is sent in response to the successful reception of a specific preamble. Only the UEs that receive an uplink grant can continue with the RAP.

As a result, the signaling capacity of an eNB is limited by the number of available preambles and by the number of uplink grants that can be sent per RAR message. This capacity, can be easily exceeded when a bulk of M2M UEs transmit their preambles in a highly synchronized manner, which is a typical behavior in mMTC applications

that leads to severe congestion. Congestion caused by mMTC applications is a serious problem, as the rapid increase in the number of deployed M2M UEs will undoubtedly increase the frequency and severity of congestion in the near future.

Recent efforts to support mMTC in LTE-A have led to the development of the narrowband Internet of Things (NB-IoT) standard, presented in release 13 of the 3GPP specifications [7]. NB-IoT is a low-power wide-area (LPWA) implementation at the eNBs that aims to support mMTC by providing great power efficiency, low bandwidth utilization, and enhanced coverage at a reduced hardware cost. As such, NB-IoT devices are expected to remain active for up to ten years without the need of battery replacements and to communicate at a distance of up to ten kilometers from the eNBs [7]. Nevertheless, the RAP in traditional LTE-A, in NB-IoT, and in 5G is mostly similar, with only a few minor exceptions that have a minor impact on performance. Therefore, it is only natural that the development of efficient access control schemes is a hot research topic [12, 35, 37, 58, 90, 106, 109].

Among the numerous access control schemes that have been proposed in the literature, the access class barring access class barring (ACB) scheme is one of the most promising; hence it has been included in the 3GPP Radio Resource Control (RRC) specification [10]. The ACB scheme redistributes the UE access attempts through time. For this, the eNB may force the UEs to randomly delay the beginning of RAP according to the barring parameters: barring rate and mean barring time. By doing this, it may be effective to relieve sporadic and short (in the order of a few seconds) periods of congestion. This behavior goes in line with the bursty traffic behavior of mMTC applications [1, 109]. The ACB scheme is explained in detail in Section 2.3.

Throughout these studies, it was identified that the behavior of ACB is oftentimes misinterpreted in the literature. That is, we have observed that some studies that evaluate the efficiency of the ACB scheme, such as Lin *et al.* [71] assume that the time the UE accesses are delayed is fixed, whereas the 3GPP technical specifications state that this parameter is selected randomly at each barring check. A barring check is the process by which the UE determines its barring status, please refer to Section 2.3 for specific details on the ACB scheme [5, 10]. Our studies are one of the few that evaluate

the ACB performance with a randomly selected barring time.

This chapter presents a thorough performance analysis of the RA in cellular networks, which includes both, the RAP and the ACB scheme, in mMTC scenarios. For this, typical configurations and the timing of the LTE-A RAP are assumed, but our results can be easily extended to NB-IoT and 5G just by including minor modifications on some configuration parameters. Throughout this chapter, we focus on the performance analysis of the ACB scheme with a static configuration. That is, the barring parameters remain constant throughout the whole period in which UE accesses occur. That is, throughout the whole distribution period. On the other hand, possible methods to adapt the barring parameters to the signaling traffic intensity in real time are discussed in Chapter 4. Building on this, the main contributions of this chapter are as follows.

1. The analysis of the steady-state capacity of the RAP.
2. The identification of the combinations of configuration parameters that enhance the success probability in mMTC scenarios.
3. The comparison of the key performance indicators (KPIs) obtained for two different backoff (i.e., time elapsed between a failed preamble transmission and the next preamble transmission) implementations at the UE side:
   (a) a uniform backoff (as stated in the medium access control (MAC) specification [9]).
   (b) an exponential backoff, where the backoff time of each UE depends on the number of transmissions attempted previously.
4. A thorough analysis of the ACB scheme that allows to identify the optimal parameter configuration under the most congested scenario suggested by the 3GPP [1].

The rest of the chapter is organized as follows. Section 2.2 presents a review of the literature on mMTC in cellular networks. Next, Section 2.3 describes the ACB scheme and the RAP in detail; including the analysis of the capacity of the RAP. Section 2.4

presents the methodology and the results derived from the performance analysis of the RAP by simulation. The potential benefits of the ACB scheme and its optimal configuration are also included in this section. Finally, conclusions are presented in Section 2.5.

2.2 Related work

The RAP of 3GPP cellular networks is complex as it comprises a four-message handshake and involves two main physical channels; each of these channels presents different limitations. It is this same complexity that makes the RAP inefficient under mMTC scenarios, but also difficult to evaluate properly. For instance, it involves numerous configuration parameters and each of these can take several values that have a deep impact on performance.

The 3GPP has provided a list of typical configuration parameters and the resulting KPI, along with some recommendations which serve as initial guidelines for performance analysis of the RAP [1]. Several studies have concluded that the RAP cannot handle a large number of UE accesses efficiently, especially when the UE accesses are highly synchronized [24, 37, 58, 79, 109]. Nevertheless, some studies simply adopt the default configuration provided by the 3GPP and overlook other possible combinations of parameters. For example, Wei *et al.* [109] present a thorough mathematical analysis of the RAP that, due to its high accuracy, will be used as a benchmark to our analytical model in Chapter 3. However, their work focuses on the analytical modeling rather than on the performance evaluation of the RAP under mMTC scenarios. Consequently, only the typical configuration is evaluated.

Lin *et al.* defined the physical RACH (PRACH) capacity and were one of the firsts to investigate the benefits of the ACB scheme with static parameters [71]. They also proposed one of the first dynamic schemes to configure the ACB scheme. However, the capacity of the physical downlink control channel (PDCCH), where the second message of the RAP is transmitted, was not considered and only a few combinations of barring parameters were tested. Furthermore, the time the UE accesses are delayed

due to ACB was considered to be fixed [1], whereas this time is calculated randomly at each barring check [5, 10]. In addition, performance was only evaluated under the most typical configuration.

De Andrade *et al.* evaluated the performance of the RAP by means of a commercial simulator and compared the benefits of several schemes to relieve congestion [34]. They found that: 1) the ACB scheme is one of the most effective solutions to congestion, even when barring parameters are fixed; 2) it is difficult to correctly adapt the barring parameters to the traffic intensity in real time; and 3) an exponential backoff may be helpful to decrease congestion. However, two main drawbacks were observed in this study. The first one is that the authors assumed that preambles transmitted by multiple UEs are decoded by the eNB, whereas the 3GPP states that these are opposite outcome occurs. Possible scenarios for these two possible outcomes, their implications, and their impact on performance are discussed in detail in Chapter 3. As it will be seen, this assumption greatly affects the performance of the RAP. The second drawback is that numerous schemes were investigated, but none of these was thoroughly described or evaluated. For example, only one combination of parameters was selected for each of them. Building on this, a thorough performance analysis of the RAP and that includes the ACB scheme as defined in the technical specifications is needed.

2.3 Random access in LTE-A

This section provides a detailed description of the contention-based RA in cellular networks, which includes the ACB scheme and the RAP itself. The capacity of the channels involved in the RAP (i.e., the capacity of the RAP) is also derived.

Through Chapters 2 to 4, we assume the ACB scheme is exclusively implemented as described in the technical specifications [5, 10]. That is, no other access control scheme is considered. On the other hand, we consider the complete four-message handshake performed in the contention-based RAP. These are now described in detail.

In order to switch from idle to connected mode, the UEs must first acquire the network configuration parameters; these are broadcast by the eNB through System

[1] This was confirmed by simulation and by our analytical model.

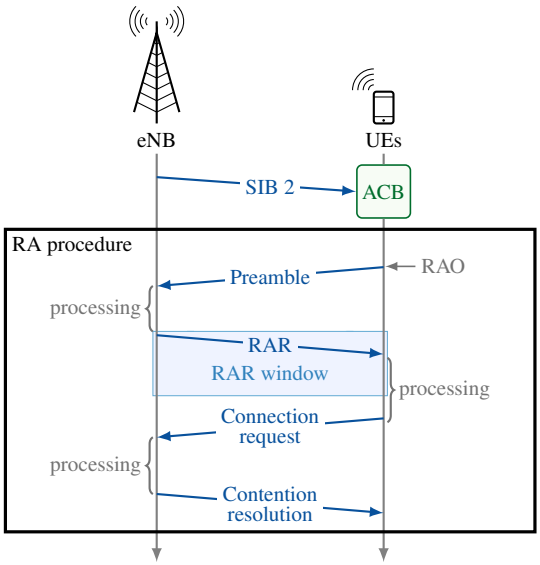


Figure 2.1: Contention-based RA in cellular networks. UEs are first subject to the ACB scheme; they perform the RAP afterwards.

Information Blocks (SIBs). Concretely, the basic network configuration is included in the Master Information Block (MIB), *PRACH-ConfigSIB* and in SIB 1 and SIB 2 [10].

The network operates in a time-slotted channel in which the minimum unit for scheduling is the subframe, with duration of $t_s = 1$ ms. The MIB includes the available carrier bandwidth, whereas the *PRACH-ConfigSIB* includes the parameter *prachConfigIndex*, which defines the period of RAOs. That is, the number of subframes elapsed between two consecutive RAOs. The SIB 1 carries the scheduling information for the remaining SIBs such as the period of SIB transmissions. Finally, SIB 2 includes, among others, the barring parameters [10]. Once the UEs have acquired the necessary information, they continue to the ACB scheme and, finally, the RAP. Fig. 2.1 briefly describes the contention-based RA.

Upon arrival, UEs are subject to the ACB scheme and are divided into access

classes (ACs) 0 to 15 according to their traffic characteristics. Each UE belongs to one out of the ten normal ACs (from ACs 0 to 9), and can also belong to one or more out of the high-priority classes. For instance, AC 10 is for emergency calls and ACs 11 to 15 are special ACs.

Please recall that RAOs are time-frequency resources in which preamble transmissions (i.e., first step of the RAP) are allowed. Next, let i be the time index that defines the number of RAOs elapsed since the beginning of an observation of the system and up to a given point in time. Also let j be the time index that defines the number of SIB 2 transmissions performed up to the ith RAO. Building on this, the jth SIB 2 transmission includes the barring rate $p_{\text{acb}}(j)$ and the mean barring time $t_{\text{acb}}(j)$ that are applied to all ACs 0 to 9, and to one or more of the ACs 10 to 15 until the $(j+1)$th SIB 2 transmission. The list of the high-priority categories that are subject to the ACB scheme is also included in the SIB 2 [10].

The UEs subject to the ACB scheme must perform a barring check before initiating the RAP (i.e., before the transmission of their first preamble) as described in Algorithm 1 [5, 10, 53]. On the other hand, the UEs that succeed in a barring check are no longer subject to the ACB scheme and proceed to perform the RAP as follows.

Preamble (*Msg1*): Let r be the number of available preambles for the contention-based RAP. Each UE randomly selects one out of the r available preambles and sends it toward the eNB in the next RAO through the PRACH. Preambles in LTE-A are orthogonal (i.e., Zadoff-Chu) sequences, whereas in NB-IoT these are orthogonal single-tone frequency-hopping patterns. Therefore, the number of available preambles in traditional LTE-A is limited by the characteristics of Zadoff-Chu sequences to a maximum of $r = 64$. On the other hand, the number of available preambles in NB-IoT is limited by the preamble and system bandwidth to a maximum of $r = 48$ [10].

Due to the orthogonality of preambles, multiple UEs can access the eNB in the same RAO if they select different preambles. That is, the eNB decodes the preambles transmitted with sufficient power by exactly one UE in each RAO. On the other hand, a collision occurs when multiple UEs transmit the same preamble simultaneously.

RAR (*Msg2*): The eNB computes an identifier for each successfully decoded preamble and sends the RAR message through the PDCCH. It includes, among other

Algorithm 1 ACB scheme.

1: **repeat**
2: Select the mean barring time $t_{\text{acb}}(j)$ and barring rate $p_{\text{acb}}(j)$ broadcast by the eNB in the jth SIB 2.
3: Generate $\text{U}[0, 1) \equiv$ a random number with uniform distribution between 0 and 1.
4: **if** $\text{U}[0, 1) \leq p_{\text{acb}}(j)$ **then**
5: Initiate the RAP.
6: **else**
7: Generate a new $\text{U}[0, 1)$.
8: Select the barring time as

$$t_w = (0.7 + 0.6\ \text{U}[0, 1))\, t_{\text{acb}}(j). \tag{2.1}$$

9: Wait for t_w.
10: **end if**
11: **until** the RAP is initiated.

data, uplink grants for the transmission of *Msg3* in predefined time-frequency resources. There can be up to one RAR message in each subframe, but it may contain several uplink grants; each of which is associated to a successfully decoded preamble.

The PDCCH resources are limited, so a maximum number of uplink grants can be sent per RAR message and only one RAR message can be sent per subframe. After preamble transmission, UEs wait for a predefined number of subframes to receive the uplink grant. This period is known as the RAR window. Hence, the number of available uplink grants per RAR window depends on the length of the RAR window and on the number of uplink grants that can be sent per subframe.

Connection request (*Msg3*): After receiving the corresponding uplink grant, the UEs adjust their uplink transmission time according to the received time alignment and

schedule the transmission of the connection request message toward the eNB through dedicated resources.

Contention resolution (*Msg4*): The eNB transmits a contention resolution message in response to each received *Msg3*. If a *Msg3* transmission fails, the eNB will not send the *Msg4* and the UE schedules a *Msg3* retransmission a few subframes later. However, if a UE does not receive *Msg4* within a predefined time window known as the contention resolution timer or within the maximum number of transmission attempts, then it declares a failure in the contention resolution and schedules a new preamble transmission. These parameters are provided by the eNB. It is important to emphasize that a failure at the contention resolution is extremely rare if common values for the two parameters are selected, and may only occur under poor wireless conditions.

There exists a maximum number of allowed preamble transmissions for each UE. This number is broadcast by the eNB through the SIB 2 [10] and, as it will be seen in Section 2.4, it plays an important roll in the performance of the network. Whenever an access attempt fails, and if the maximum number of preamble transmissions has not been reached, the UE waits for a random backoff time (determined by the backoff indicator); then randomly selects and transmits a new preamble at the next RAO. UEs perform a power ramping process to reduce the probability of subsequent preamble transmission failures due to wireless channel errors. In this process, UEs transmit the first preamble with low power; then, transmission power increases at each failed access attempt.

As described above, preambles transmitted by exactly one UE with sufficient power are decoded at the eNB. On the other hand, two possible outcomes exist when the same preamble is transmitted by multiple UEs simultaneously. In the first one, the eNB does not decode the transmitted preamble. Hence, the implicated UEs will not receive an uplink grant within the RAR window; this is the indication that a collision has occurred. In the second one, the eNB correctly decodes the transmitted preamble and may send an uplink grant in response; this uplink grant will be received by multiple UEs. Each uplink grant assigns time-frequency resources for the transmission of *Msg3*. Hence, the implicated UEs will transmit their *Msg3*s in the same dedicated resources. Therefore, the preamble collision will be detected at this point.

In this chapter we assume that the first outcome takes place whenever a preamble collision occurs. This goes in line with the 3GPP recommendations for the performance analysis of the RACH [1] and with most of the literature [20, 30, 71, 94, 109, 116]. Chapter 3 delves deep into the multiple causes for the two different outcomes mentioned above, along with the two main assumptions related to the RAP and their impact on performance.

It is important to mention that we assume the available resources and the timing of the RAP are the ones defined for traditional LTE-A. These are shown in Table 2.2 on page 25 and in Table 2.3 on page 26, respectively. But is also worth emphasizing that the RAP in NB-IoT and 5G are greatly similar to that in LTE-A with a few minor exceptions [4]. Concretely, the single difference between NB-IoT and LTE-A that may have an impact on performance is that up to three coverage enhancement (CE) levels can be defined in NB-IoT (CE levels zero, one, and two). The CE level of a given UE defines the number of preamble repetitions to be performed one after another per each access attempt. That is, only one repetition performed at CE level zero and the number of preamble repetitions increases with the CE level. Preamble repetitions are meant to reduce the probability of an access failure due to wireless channel errors [44, 50].

Specifically, every UE in NB-IoT belongs to CE level zero unless the quality of the measured reference signals sent by the eNB is poor due to an unfavorable wireless environment, or the UE has reached the maximum number of access attempts successfully. In the latter case the UE increases the CE level. Building on this, the ratio of UEs in CE level zero to UEs in CE levels one and two is expected to be considerably large when no congestion has occurred. Furthermore, the preambles assigned to each CE may be different. Building on this, the contributions presented in chapters 2, 3, and 4 can be easily and successfully applied to the access control of the UEs in CE level zero as these contribute the most to congestion in the RACH given that different preambles have been assigned to each CE level.

2.3.1 Capacity of the RAP

As it will be showcased throughout this section, the capacity of the channels involved in the RAP is determined by two main parameters. The first one is the number of

available preambles for contention-based RA, denoted by r. The second one is the number of available *uplink grants* per RAR window, denoted by g. In this section we first evaluate the PRACH capacity per RAO, which then is immediately extended to the capacity of the RAP, including both the PRACH and the PDCCH. With this information, the capacity of the RAP given in successful accesses per second can be easily calculated.

As mentioned above, preambles are constructed using Zadoff-Chu sequences [8]. These are orthogonal sequences possess great periodic correlation, which allows for an extremely fast calculation of their correlation [31]. Nevertheless, Zadoff-Chu sequences are difficult to generate in real time and require large amounts of memory for their storage [74]. On the other hand, preambles in NB-IoT are single-tone frequency hopping patterns, where frequency hopping is pseudo aleatory. Therefore, a preamble in NB-IoT is defined by the initial tone selected by the UEs. Building on this, the RACH of both LTE-A and NB-IoT resembles a multichannel slotted ALOHA network access protocol in which r corresponds to the number of available channels. Time slots correspond to the minimum time unit for scheduling, which in LTE-A is the subframe, with a fixed duration of $t_s = 1$ ms.

Let S be the random variable (RV) that defines the number of successful preambles at an arbitrary RAO. That is, preambles selected by exactly one UE at a given RAO. The state space of S is the number of successes $\{s \in \mathbb{N} \mid s \leq r\}$. Also, let $n(i)$ be the number of contending UEs at the ith RAO. That is, the total number of preamble transmissions at the ith RAO. The expected value of S at the ith RAO is given as

$$\mathbb{E}[S] = n(i)\left(1 - \frac{1}{r}\right)^{n(i)-1} \tag{2.2}$$

Fig. 2.2 shows $\mathbb{E}[S]$ for several values of r and $n(i)$. As it can be seen, $\mathbb{E}[S]$ is an increasing function for low values of $n(i)$ and presents its global maximum when $n(i) \approx r$. Then it becomes a decreasing function.

Based on this behavior, the following definition for the capacity of the PRACH was formulated in [71].

Definition 2.3.1. *The PRACH capacity* $\mathrm{C}(r)$ *is given as the maximum achievable* $\mathbb{E}[S]$ *for any* $n(i) \in \mathbb{R}$ *and for a given* r.

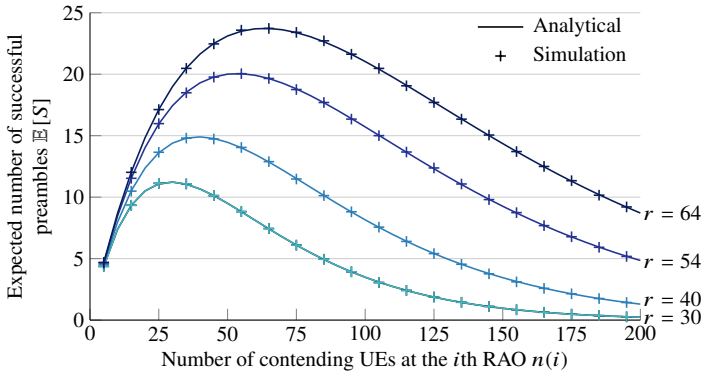


Figure 2.2: Expected number of successful preambles $\mathbb{E}[S]$ given r available preambles and $n(i)$ contending UEs [71, Fig. 3].

Lemma 2.1. *The value of* $n(i) \in \mathbb{R}$ *that maximizes* $\mathbb{E}[S]$ *as formulated in* (2.2) *is* $n^*(i) = \lceil \log(r/[r-1]) \rceil^{-1}$. *Therefore, the capacity of the PRACH is given as*

$$\mathrm{C}(r) = \max_{n(i)} \mathbb{E}[S] = \left\lceil \log\left(\frac{r}{r-1}\right)\right\rceil^{-1} \left(1 - \frac{1}{r}\right)^{\left\lceil \log\left(\frac{r}{r-1}\right)\right\rceil^{-1}-1}. \tag{2.3}$$

Proof. The value $n^*(i)$ can be easily obtained by means of the first derivative test as follows.

$$\frac{\partial\, \mathbb{E}[S]}{\partial n(i)} = \left(1 - \frac{1}{r}\right)^{n(i)-1} \left[1 + n(i) \log\left(1 - \frac{1}{r}\right)\right] = 0 \tag{2.4}$$

which gives

$$n^*(i) = -\frac{1}{\log\left(\frac{r-1}{r}\right)} = \left\lceil \log\left(\frac{r}{r-1}\right)\right\rceil^{-1}. \tag{2.5}$$

This concludes the proof. □

We observed in [95] that the PRACH capacity can be closely approximated by the

following simple formulations.

$$\mathrm{C}(r) \approx r\left(1-\frac{1}{r}\right)^{r-1} \tag{2.6}$$

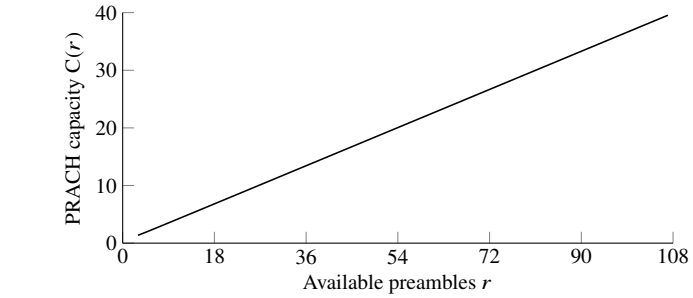
$$> \frac{r}{e} \tag{2.7}$$

where e is Euler's number. In particular, both (2.6) and (2.7) are lower bounds of (2.3); these are derived in Appendix B.1.

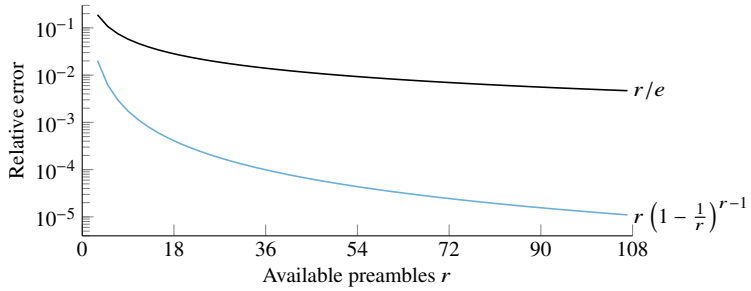
Fig. 2.3 shows $\mathrm{C}(r)$ as in (2.3) and the relative error of lower bounds (2.6) and (2.7). Clearly, the error of approximation (2.6) is negligible and can be directly used instead of (2.3), but also that of approximation (2.7) is relatively low for practical values of r. Typical values of r can be inferred from the fact that a total of 64 and 48 preambles exist in LTE-A and NB-IoT, respectively. For instance, $r = 54$ is the most typical value in LTE-A [1], whereas $r = 30$ for NB-IoT seems natural. For these values of r we have $\mathrm{C}(54) = 20.05$ and $\mathrm{C}(30) = 11.22$, whereas for $r = 30$ the relative error with (2.6) is below $1.7 \cdot 10^{-2}$ and with (2.7) is below $1.5 \cdot 10^{-4}$.

It will be showcased in Fig. 2.4 on page 27 that $\mathrm{C}(r)$ approximately coincides with the maximum number of stationary UE arrivals per RAO that the PRACH can handle efficiently. That is, when the whole RAP is performed and no limitations on the PDCCH are considered. For instance, please assume the most typical value of $r = 54$ is selected, which gives a theoretical capacity of $\mathrm{C}(54) = 20.05$ successful preambles. In a typical PRACH configuration RAOs occur once every $t_{\mathrm{rao}} = 5$ ms. Since the subframe duration is t_s1 ms, RAOs occur once every 5 subframes. Therefore, the maximum number of stationary UE arrivals per second (i.e., stationary signaling traffic load) that the most typical PRACH configuration can handle efficiently is approximately $\mathrm{C}(r)/t_{\mathrm{rao}} = 4010$. In other words, 4010 successful accesses per second can be achieved when the full capacity of the PRACH is utilized. Nevertheless, the number of available preambles is not the only parameter that limits the capacity of the RAP and, as it will be observed throughout this thesis, the capacity of the PDCCH oftentimes has a greater impact on the performance of the RAP.

Specifically, the capacity of the PDCCH is measured in terms of control channel elements (CCEs), and is fixed to 16 CCEs per subframe. Both, RAR and contention



(a)



(b)

Figure 2.3: (a) PRACH capacity defined as in [71] and (b) relative error of approximations (2.6) and (2.7).

resolution messages (i.e., *Msg2* and *Msg4*, respectively) are transmitted through the PDCCH, hence resources are shared among these two messages. The size of each uplink grant included in *Msg2* and of *Msg4* is four CCEs. At each subframe, one RAR message can be sent and at least four CCEs are reserved for a *Msg4* transmission [79]. Therefore, the maximum number of uplink grants can be sent per ms is $f_g = 3$. As mentioned at the beginning of this section, the UEs wait for a predefined number of subframes after preamble transmission to receive the uplink grant; this is called the RAR window and its duration under the most typical configuration is $t_{\text{rar}} = 5$ ms (i.e., equal to the period of RAOs). As a result, the maximum number of uplink grants that can be sent within a RAR window is $g = f_g\, t_{\text{rar}}$. Building on this, we extend Definition 2.3.2 as follows

Definition 2.3.2. *Let N_G be the RV that defines the number of UEs that receive an uplink grant at a given RAR window. That is, in response to preambles transmitted in the previous RAO. Hereafter we refer to N_G as the number of assigned uplink grants. The capacity of the RAP is defined as the global maximum of $\mathbb{E}[N_G]$ for any $n(i)$, given no wireless channel errors occur during preamble nor uplink grant transmissions. Hence, the capacity of the RAP under the assumption that the eNB only decodes preambles transmitted by exactly one UE is simply given as*

$$\mathrm{C}(r, g) = \max_{n(i)} \mathbb{E}[N_G] = \min\{\mathrm{C}(r), g\}. \tag{2.8}$$

2.4 Performance analysis of RA in cellular networks

In this section, we first describe the methodology for the performance analysis of the RA in cellular networks under mMTC scenarios. Next, we investigate the relation between the capacity of the PRACH alone and the maximum static number of arrivals that this channel can handle efficiently. Then, we investigate the performance of the RAP under different configurations of the PRACH, along with the potential benefits of implementing an exponential backoff instead of the traditional uniform backoff. Finally, we evaluate the efficacy of the ACB scheme and identify its optimal configuration.

2.4.1 Methodology

The 3GPP has condense the enormous collection of variables and possible mMTC scenarios in two different traffic models and also has provided five KPIs to assess the performance of the RAP. The characteristics of the traffic model (TM) 1 and of the TM 2 are shown in Table 2.1 [1]. Both traffic models correspond to mMTC scenarios that are expected to occur in the near future as their characteristics are based on observations performed in a highly dense urban area. However, the main difference between them is that the TM 2 corresponds to a highly synchronized scenario, whereas UE arrivals under the TM 1 are uniformly spread across the whole distribution period.

Table 2.1: Characteristics of the different traffic models defined by the 3GPP for the performance evaluation of the RAP [1].

Parameter	TM 1	TM 2
Number of M2M UEs n	{1000, 3000, 5000, 10 000, 30 000}	{1000, 3000, 5000, 10 000, 30 000}
Distribution period t_{dist} (s)	60	10
Distribution over t_{dist}	Uniform	Beta(3, 4)

It has been observed that TM 2 causes severe congestion when the most typical PRACH configuration is selected [1]. Therefore, we assume the TM 2 represents the default behavior under mMTC scenarios and is used throughout the vast majority of results presented in this section. In particular, we select $n = 30\,000$ UE arrivals that, according to TM 2, follow a Beta(3, 4) distribution over 10 s.

The KPIs proposed by the 3GPP for the performance analysis of the RA are defined as follows.

1. Success probability P_s: Ratio of successful to total UEs. To calculate this parameter, let $s(i)$ be the number of UEs that successfully complete the RAP at the ith RAO. Then, P_s is simply given as

$$P_s = \frac{1}{n} \sum_{i=0}^{i_{\max}} s(i) \tag{2.9}$$

   Throughout Chapters 2 to 4, P_s is considered the most important KPI and assume performance is adequate only if $P_s \geq 0.95$.

2. Access delay D: RV that defines the time elapsed between the arrival of a UE and the successful completion of the RAP, given in seconds. We assess D in terms of the 10th, 50th, and 95th percentiles denoted as D_{10}, D_{50}, and D_{95}, respectively. That is, the delay of ϕ percent of the UEs that successfully complete the RAP is D_ϕ s or less.

3. Preamble transmissions K: RV that defines the number of preamble transmissions performed by the UEs that successfully complete the RAP. As such, the

transmissions performed by UEs that fail the RAP are not considered. We assess K in terms of the expected value $\mathbb{E}[K]$.

4. Collision probability P_c: Ratio of collided to total number of available preambles in the access period. For this, let $c(i)$ be the number of collided preambles at the ith RAO (i.e., preambles transmitted by multiple UEs). Also let $i_{\max}$ be the last RAO of the access period. Then it follows that

$$P_c = \sum_{i=0}^{i_{\max}} \frac{c(i)}{r}. \tag{2.10}$$

5. Number of contending UEs per RAO: Equivalent to the total number of preamble transmissions in a RAO. This KPI serves as an indicator of the levels of congestion of the RA channels.

The first three KPIs listed above will be the base to assess the performance of the RAP. The reasons to overlook the collision probability and the number of contending UEs are:

1. The collision probability highly depends on the period that is taken into account for its calculation. That is, $\{i \in \mathbb{N} \mid i \leq i_{\max}\}$, but $i_{\max}$ is not strictly defined. Hence, results are not easily comparable with those obtained in the literature, for example, with those included in [1]. This problem becomes more evident when the ACB scheme is introduced.

2. The number of contending UEs per RAO directly impacts the number of successful and collided preambles. Hence, this KPI is directly reflected on P_s, but the latter provides more information on the actual performance of the RAP. Building on this, we use P_s as the primary KPI and employ the number of contending UEs per RAO exclusively for illustration purposes.

A static implementation of the ACB scheme is considered throughout this section and also throughout Chapter 3. This means that the barring parameters remain constant throughout the operation of the network or, at least, throughout a whole observation of it. In our case, each observation of the network begins at the first RAO in the

Table 2.2: Default parameters for simulations.

Parameter	Setting
Available preambles	$r = 54$
Subframe length	$t_{\mathrm{s}} = 1$ ms
Period of RAOs	$t_{\mathrm{rao}} = 5$ ms
RAR window length	$t_{\mathrm{rar}} = 5$ ms
Available uplink grants per RAR window	$g = 15$
Maximum number of preamble transmissions	$k_{\max} = 10$
Backoff indicator	$b_{\max} = 20$ ms
Error probability for the kth preamble transmission	$\Pr[\mathcal{E}_k] = 1/e^k$
Maximum number of *Msg3* and *Msg4* transmissions	5
Error probability for *Msg3* and *Msg4* transmissions	0.1

distribution period and ends when every M2M UE has concluded the RAP. This allows us to simplify notation by defining $p_{\mathrm{acb}} = p_{\mathrm{acb}}(j)$ and $t_{\mathrm{acb}} = t_{\mathrm{acb}}(j)$ for all $j \in \mathbb{Z}_+$.

Results presented in this section were obtained by means of a C-based simulator that closely replicates the ACB scheme, the arrival process of the UEs, and the RAP as described in the specifications [5, 10]. This simulator was developed in the early stages of the PhD program. In each simulation, the n M2M arrivals are scheduled within t_{dist}, which begins at the zeroth RAO (i.e., $i = 0$). Each simulation ends when every UE has terminated the RAP. The number of simulation runs is set to the smallest number that ensures that all the cumulative KPIs obtained up to the last simulation differ from those obtained up to the previous simulation by less than 0.01 percent. The default configuration of the PRACH and PDCCH, along with the model for the quality of the wireless channel are shown in Table 2.2. These were suggested by the 3GPP for the performance evaluation of the RA in LTE-A and are used throughout this section unless otherwise stated.

Furthermore, the timing parameters of the RAP, also defined by the 3GPP, are shown in Table 2.3 [3, Table 16.2.1-1].

Table 2.3: Timing of the four-message handshake in the LTE-A RAP [3, Table 16.2.1-1].

Message	Time (ms)
Preamble processing delay	2
Uplink grant processing delay	2
Connection request processing delay	4
Connection request round-trip time (RTT)	8
Contention resolution RTT	5

2.4.2 Results

As a starting point for the performance analysis of the RA in cellular networks, we showcase the relation between the PRACH capacity $\mathrm{C}(r)$ and the maximum number of stationary UE arrivals that the PRACH can handle efficiently. For this, let $x(i)$ be the number of UE arrivals at the ith RAO. As no access control scheme is implemented, $x(i)$ is also the number of UEs that transmit their first preamble at ith RAO. We use our simulator to replicate the complete RAP and to generate a stationary distribution of $x(i) \in \{1, 2, \ldots, 40\}$ UEs per RAO during a period that is long enough to consider the system reaches a steady state. We eliminate the effect of the PDCCH capacity by setting $g = r$. This allows us to evaluate the PRACH alone.

Fig. 2.4 shows P_s as a function of $x(i)$ for different values of r. As expected, $P_s \approx 1$ for low values of $x(i)$, but then an abrupt drop occurs at approximately $x(i) = \mathrm{C}(r)$, as calculated by (2.3). For example, $P_s \approx 1$ until $x(i) \approx 20$ for $r = 54$, which gives $\mathrm{C}(54) = 20.05$. Therefore, Fig. 2.4 demonstrates an important fact: if the PRACH were the only limiting factor of the RAP, $\mathrm{C}(r)$ would be a close upper bound to the stationary number of accesses per RAO that the PRACH can handle efficiently. As stated above, during these tests we merely evaluated the impact of the PRACH capacity, but similar conclusions can be drawn when $g < \mathrm{C}(r)$. To support this claim, we proceed to evaluate the performance of the RAP under the TM 2 with $n = 30\,000$ and with the default configuration shown in Table 2.2.

Fig. 2.5 shows the average number of UE arrivals, contending UEs, collided pream-

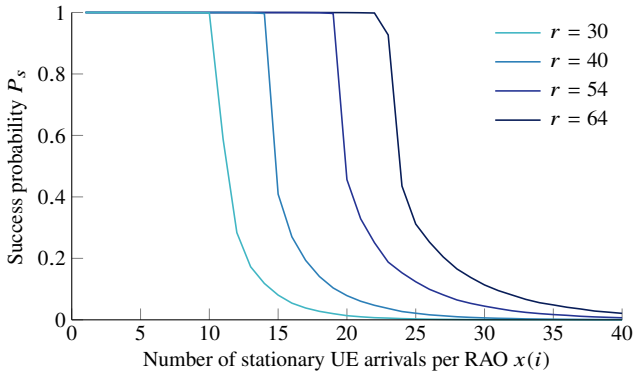


Figure 2.4: Success probability P_s given $x(i)$ stationary UE arrivals per RAO.

bles, and successful accesses per RAO under the TM 2. It can be easily observed in Fig. 2.5 that the average number of contending UEs is far greater than the UE arrivals; this effect is due to the high congestion that builds up when the capacity of the RAP is exceeded. As it can be seen from Fig. 2.5, under these conditions the average number of UE arrivals exceeds $C(54, 15) = 15$ from the 343th until the 1329th RAO, which results in a congestion period of almost 5 s (i.e, 986 RAOs). The peak of congestion occurs at exactly the 800th RAO, in which slightly more than 300 contending UEs (i.e., preamble transmissions) are observed on average. This point exactly coincides with the global maximum of the average number of UE arrivals. Due to the exceedingly large number of contending UEs, the average number of successful accesses is extremely low during this period and a poor $P_s = 0.313$ is obtained. It is worth noting that our results closely match those obtained by the 3GPP under this same conditions [1]; this validates the correct operation of our simulator.

After these results were obtained, we focused on enhancing the performance of the RAP by manipulating its configuration parameters and by implementing a different backoff as the one defined in the 3GPP specifications [10]. That is, without incorporating the ACB or any other access control scheme. As Fig. 2.6 and Fig. 2.7 illustrate, merely manipulating the configuration parameters or implementing a different backoff is not sufficient to prevent congestion under highly synchronized mMTC scenarios.

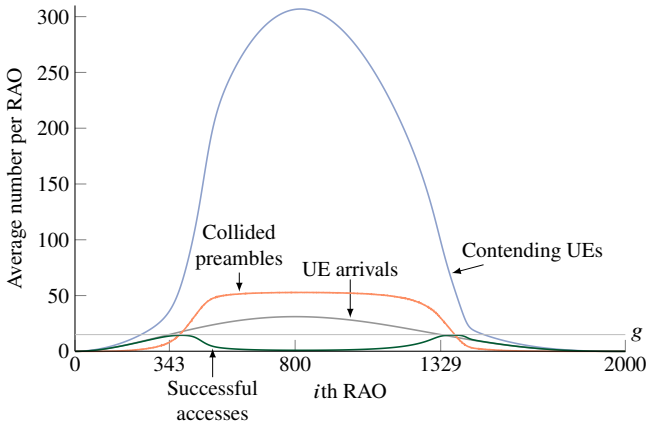


Figure 2.5: Average number of UE arrivals, contending UEs, collided preambles, and successful accesses per RAO under the TM 2 with $n = 30\,000$ M2M UEs.

Nevertheless, simple adjustments to some parameters such as the maximum number of preamble transmissions or implementing an exponential backoff can lead to an increase in P_s of around 30 percent when compared to the most typical configuration. In these tests, H2H traffic was injected and H2H UE arrivals are uniformly distributed over time at a rate of $\lambda = 1$ arrivals per second.

We begin our analysis of the impact on performance of configuration parameters with the most intuitive one: the number of available preambles r. That is, we assume the number of available uplink grants is not a limitation (i.e., $g = r$) and investigate the number of preambles needed to achieve $P_s \geq 0.95$ under the TM 2.

For this, please imagine r has no upper limit. We are set to find the minimum value of r that results in $P_s \geq 0.95$ under the TM 2; this is said to be the optimal value of r, denoted as r^*. The most intuitive approach to find r^* in a complex system is by following a brute force approach, in which possible values of r are tested until r^* is found. Naturally, the range of possible values of r can be reduced by eliminating low values for which we are sure we will obtain $P_s < 0.95$. For this, let X_i be the RV that defines the number of UE arrivals at the ith RAO; hence $\{X_i\}_{i \in \mathbb{N}}$ is a stochastic process. We want to know the maximum expected value of X_i and the RAO in which this occurs

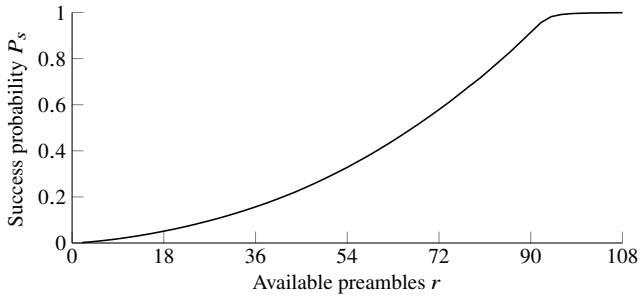


Figure 2.6: Success probability, P_s of M2M UEs only given r available preambles. M2M arrivals follow the TM 2 and $\lambda = 1$ H2H arrivals per second occur.

i^*. As mentioned above, by running a large number of simulations, we observed that $i^* = 800$, for which $\mathbb{E}[X_{i^*}] = 31.104$ arrivals occur when $n = 30\,000$ UEs follow TM 2. These values will be confirmed in Chapter 3.3 by our analytical model. From there, the first test value r^* can be obtained by means of (2.7) as follows.

$$r^* = \lceil e\ \mathbb{E}[X_{i^*}] \rceil \tag{2.11}$$

which gives $r^* = 85$.

Fig. 2.6 shows P_s for $r = \{1, 2, \ldots, 108\}$ and it can be seen that the achieved P_s with r^* is considerably high; concretely, $P_s \approx 0.8$. However, $P_s \geq 0.95$ only if $r \geq 92$. Clearly, such a dramatic increase in r is not possible in LTE-A nor in NB-IoT due to the construction of preambles. In other words, it is not feasible to incorporate such a high number of orthogonal Zadoff-Chu sequences nor subcarriers to these systems. Therefore, the impact on performance of a different configuration parameter was investigated: the maximum number of preamble transmissions $k_{\max}$.

Parameter $k_{\max}$ is signaled by the eNB through the SIB 2 [10]. Naturally, this parameter highly impacts the number of contending UEs during periods of congestion. On the other hand, few collisions occur under light traffic load scenarios such as TM 1, and the power ramping process makes it extremely rare that access failures occur due to wireless channel errors during preamble transmission. To proof this claim, let $\Pr[\mathcal{E}_k]$ be the probability that the kth preamble transmitted by a UE is successful

(i.e., exclusively selected by this UE) but lost due to a wireless channel error. Then, the probability that a given UE transmits $k_{\max}$ successful preambles without success according to the 3GPP [1] (see Table 2.2) is given as

$$\Pr\left[\bigcap_{k=1}^{k_{\max}} \mathcal{E}_k\right] = \prod_{k=1}^{k_{\max}} \frac{1}{e^k} = \frac{1}{e^{\frac{k_{\max}(k_{\max}+1)}{2}}}. \tag{2.12}$$

In other words, (2.12) defines the probability of an access failure in scenarios in which there are no other contending UEs. Still, (2.12) approximates this probability under scenarios with a relatively low traffic load, for example, TM 1.

Building on this, reducing $k_{\max}$ would have a minor impact on the probability of access failure under scenarios with a low traffic load. On the other hand, it would undoubtedly reduce the number of contending UEs during congestion and, as a consequence, increase P_s. Please observe that increasing $k_{\max}$ would have the opposite effect on the number of contending UEs and P_s under congested scenarios. Fig. 2.7a shows the achieved P_s of M2M and H2H UEs for $k_{\max} = \{1, 2, \ldots, 10\}$. Naturally, the probability of an access failure due to wireless channel errors increases slightly as $k_{\max}$ decreases. Nevertheless, the maximum P_s is achieved with $k_{\max} = 3$, which is a relatively low value.

Besides, Fig. 2.7a showcases an important fact: H2H UEs always achieve a higher P_s than M2M UEs under the TM 2. The reason for this is that H2H UE arrivals are equally distributed along the distribution period. Hence, a large portion of these UE arrivals occur when no congestion is present. For instance, $t_{\text{dist}} = 10$ s and the period of congestion is less than 5 s long. On the other hand, most of the M2M UEs arrivals are highly synchronized, so most of these occur during the period of congestion, caused by this same behavior.

Fig. 2.7b illustrates the benefits of reducing $k_{\max}$ by comparing the average number of decoded preambles and successful accesses for $k_{\max} \in \{3, 10\}$. As it can be seen, the impact of reducing $k_{\max}$ is profound as the number of successful accesses closely approximates g. In other words, reducing $k_{\max}$ reduces congestion levels, and, as a consequence, most of the available resources are utilized. This maximizes the performance of the RAP, but is not sufficient to achieve the desired $P_s \geq 0.95$.

Therefore, the potential benefits of implementing a different backoff are investigated in the following.

As mentioned in Section 2.3, a uniform backoff is envisioned in the 3GPP specifications [10]. That is, whenever a collision occurs, involved UEs generate $\mathrm{U}[0, 1) \equiv$ a random number with uniform distribution and schedule the next preamble transmission after waiting for $t_b = \mathrm{U}[0, 1)\, b_{\max}$, where $b_{\max}$ is the backoff indicator provided by the eNB. The results presented up to this point demonstrate that this uniform backoff policy, or at least with its most typical value $b_{\max} = 20$ ms, is not sufficient to avoid congestion under the TM 2. Instead, an exponential backoff in the form

$$t_b = \mathrm{U}[0, 10)\, 2^{k-1} \tag{2.13}$$

may be sufficient to spread the preamble retransmission attempts and avoid congestion. As such, the exponential backoff is implemented on M2M UEs only and H2H UEs perform the traditional uniform backoff.

Fig 2.8 shows P_s as a function of the number of M2M arrivals under the TM 2 for the uniform and exponential backoff implementations. H2H traffic has also been inserted. As can be seen, the exponential backoff increases the maximum value of n that the RAP can handle efficiently by more than 2000. That is, $P_s \geq 0.95$ for $n \leq 16\,000$ with the uniform backoff and for $n \leq 18\,000$ with the exponential backoff. However, M2M UEs achieve an insufficient $P_s < 0.6$ when $n = 30\,000$ regardless of the implemented backoff. As explained above, H2H UEs achieve a higher P_s than M2M UEs.

It is important to observe that increasing the backoff time may slightly increase P_s, but it will also increase the access delay of UEs whose preamble transmission fails due to a wireless channel error even in low traffic scenarios. Because of this, it can be concluded that modifying backoff is not a practical solution to congestion; hence, an access control scheme is needed. De Andrade et al. [34] also studied the potential benefits of an exponential backoff and came to a similar conclusion despite the fact that they considered a different collision model for their study. Furthermore, combining an exponential backoff with a reduced value of $k_{\max}$ is counterintuitive and would eliminate the benefits of these approaches. Building on this, the remainder of

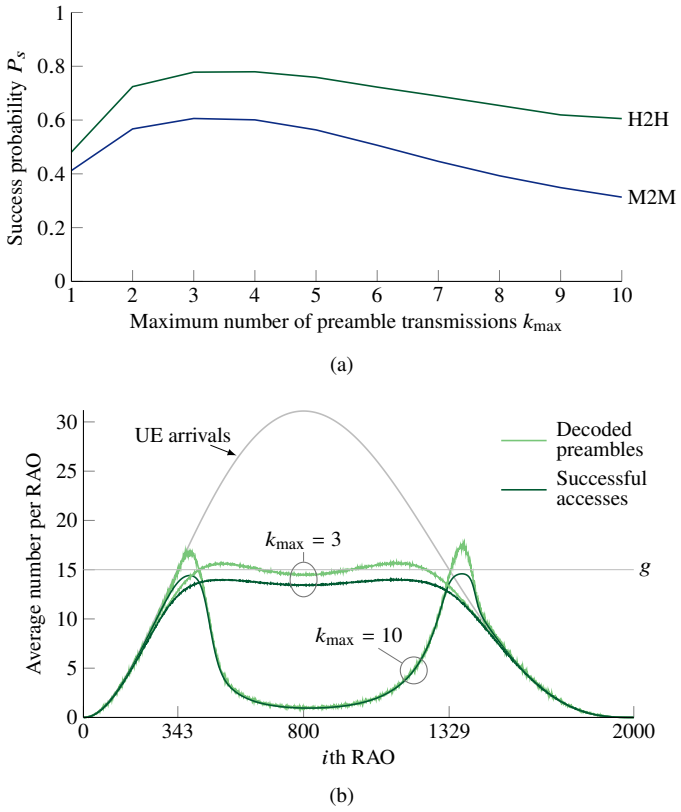


Figure 2.7: Impact of $k_{\max}$ on the performance of the RAP. (a) P_s of M2M and H2H UEs for $k_{\max} = \{1, 2, \ldots, 10\}$; M2M arrivals follow the TM 2 and $\lambda = 1$ H2H arrivals per second occur. (b) Average number of UE arrivals, decoded preambles, and successful accesses per RAO under the TM 2 given $k_{\max} \in \{3, 10\}$.

this section is dedicated to evaluate the benefits of a static implementation of the ACB scheme. That is, barring parameters, namely the mean barring time t_{acb} and the barring rate p_{acb}, remain constant during the whole observation of the system. Access class barring configurations (ACBCs) schemes, which aim to adapt the barring parameters to the signaling traffic intensity in real time are studied in Chapter 4.

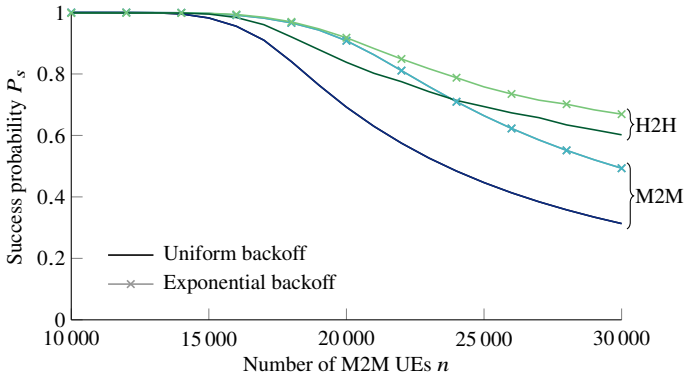


Figure 2.8: Success probability P_s of M2M and H2H UEs given n M2M UEs with uniform and exponential backoff. M2M arrivals follow the TM 2 and $\lambda = 1$ H2H arrivals per second occur.

We begin our analysis of the ACB scheme by showing the P_s achieved with different values of the barring parameters as defined in the technical specifications [10] in Fig. 2.9; H2H traffic with $\lambda = 1$ arrival per second was considered and these UEs are labeled as high priority traffic, so these are not subject to the ACB scheme. Fig. 2.9 illustrates the P_s obtained with both, the uniform and exponential backoff implementations. Clearly, the desired $P_s \geq 0.95$ can only be achieved by selecting $p_{\text{acb}} \leq 0.5$ in combination with a sufficiently long t_{acb}. The reason for this is that the maximum average number of M2M accesses is $\mathbb{E}[X_{i^*}] = 31.104$ and selecting $p_{\text{acb}} \leq 0.5$ reduces this peak to approximately g given t_{acb} is sufficiently long. Furthermore, Fig. 2.9 shows that for each $p_{\text{acb}} \geq 0.5$ there exists a maximum P_s that can be achieved, regardless of the value of t_{acb}. Once this maximum P_s is obtained, there is no reason to further increase t_{acb}.

Fig. 2.9a also shows that implementing an exponential backoff may reduce the minimum t_{acb} needed to achieve $P_s \geq 0.95$ and, again, Fig. 2.9b demonstrates that H2H UEs always achieve a higher P_s than M2M UEs. Fig. 2.9b only shows results for the uniform backoff implementation as only this is implemented in the H2H UEs and results obtained when the exponential backoff is implemented in M2M UEs are

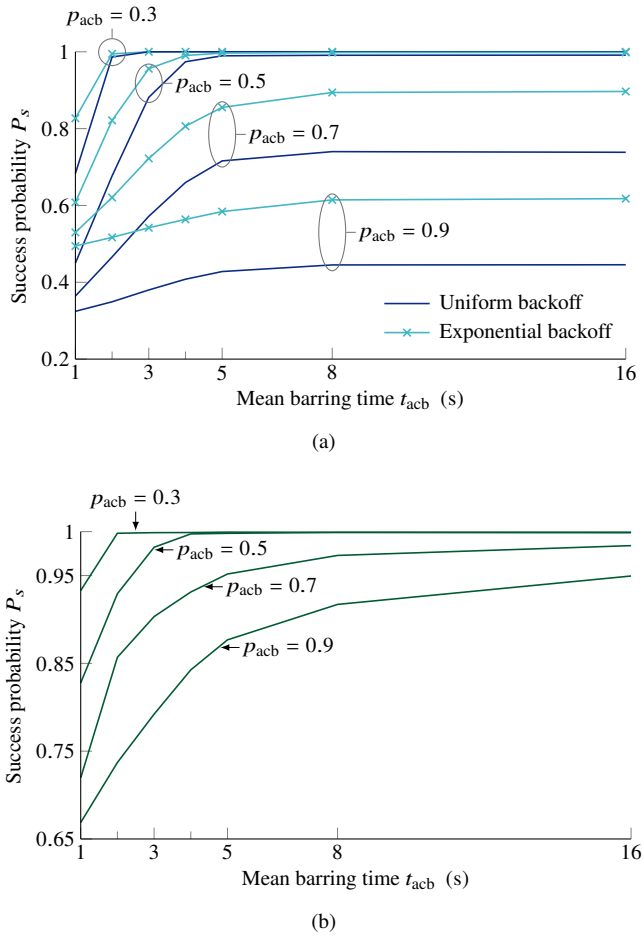


Figure 2.9: Success probability P_s of (a) M2M UEs, uniform and exponential backoff, and (b) H2H UEs under the ACB scheme.

extremely similar. Building on this, it is easily concluded that H2H traffic has no observable impact on the performance of the RA under TM 2 and that H2H UEs will always achieve a better performance than M2M UEs. Hence, we focus merely on the performance of M2M UEs.

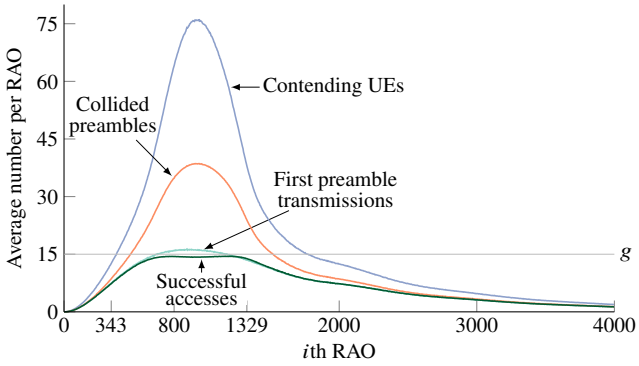


Figure 2.10: Average number of first preamble transmissions, contending UEs, collided preambles, and successful accesses per RAO under the TM 2 with $n = 30\,000$ M2M UEs given $p_{\text{acb}} = 0.5$ and $t_{\text{acb}} = 4$ s.

The reason for the efficacy of $p_{\text{acb}} \leq 0.5$ is illustrated in Fig. 2.10, where the average number of first preamble transmissions, contending UEs, collided preambles, and successful accesses per RAO are shown given $p_{\text{acb}} = 0.5$ and $t_{\text{acb}} = 4$ s. The average number of UE arrivals has been omitted as is exactly the same as that in Fig. 2.5. Clearly, this combination of barring parameters successfully prevents congestion as the number of contending UEs and collided preambles are drastically reduced when compared to those shown in Fig. 2.5. For instance, this configuration of barring parameters reduces the global maximum of the average number of contending UEs from $\approx$ 300 to $\approx$ 75 and the global maximum of the average number of collided preambles from $\approx$ 54 to $\approx$ 40. In addition, the average number of successful accesses closely follows the average number of first preamble transmissions (i.e., the number of UEs begin their RAP at each RAO). This clearly indicates the system is performing correctly and the result is a sufficiently high $P_s = 0.974$. These are promising results because they demonstrate that the ACB scheme is effective even under the TM 2, but also that it could be easily configured given the capacity of the RA channels and the global maximum number of UE arrivals in the period are known by the eNB.

Once we have identified the combinations of barring parameters that lead to $P_s \geq 0.95$, we illustrate the impact of the ACB scheme on the remaining KPIs: access

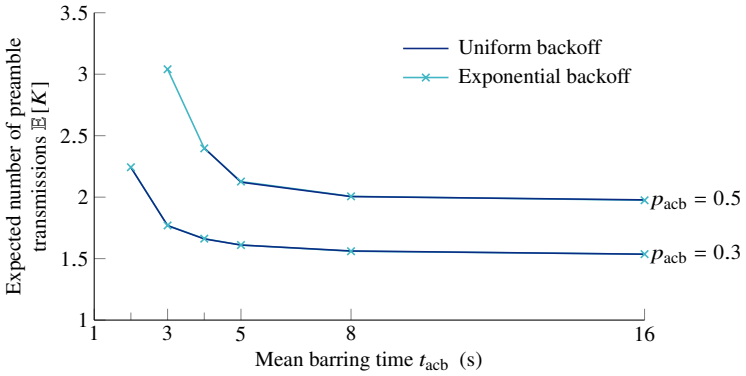


Figure 2.11: Expected number of preamble transmissions for the successful M2M UEs under the ACB scheme given $P_s \geq 0.95$.

delay and the number of preamble transmissions. We begin by showing the expected number of preamble transmissions $\mathbb{E}[K]$ for the combinations of p_{acb} and t_{acb} that lead to $P_s \geq 0.95$ in Fig. 2.11; other combinations of barring parameters have been omitted. Clearly, long values of t_{acb} decrease $\mathbb{E}[K]$ because these induce a longer delay to the UE accesses than with low values of t_{acb}; this in turn reduces the number of contending UEs and also of collided preambles. It is also clear that this is the same reason why $\mathbb{E}[K]$ decreases with p_{acb}. However, it is interesting to observe that $\mathbb{E}[K]$ is exactly the same for both the uniform and exponential backoff implementations with a given combination of barring parameters despite the small differences in P_s illustrated in Fig. 2.9a. This is another example of the profound impact that the ACB scheme has on the UE arrivals and, hence, on the performance of the RAP.

Finally, we showcase the impact of the barring parameters on the access delay D. Specifically, we focus on the 10th, 50th, and 95th percentiles of D. These are shown in Fig. 2.12 for the combinations of p_{acb} and t_{acb} that lead to $P_s \geq 0.95$. That is, the same combinations that were shown in Fig. 2.11. An important aspect that can be observed from Fig. 2.12 is that D_{10} can be up to 1000 times lower than D_{50}. Clearly, UEs with a delay $\leq D_{10}$ are the ones who succeed in the first barring check and at the first preamble transmission. This conclusion can be drawn just by comparing most

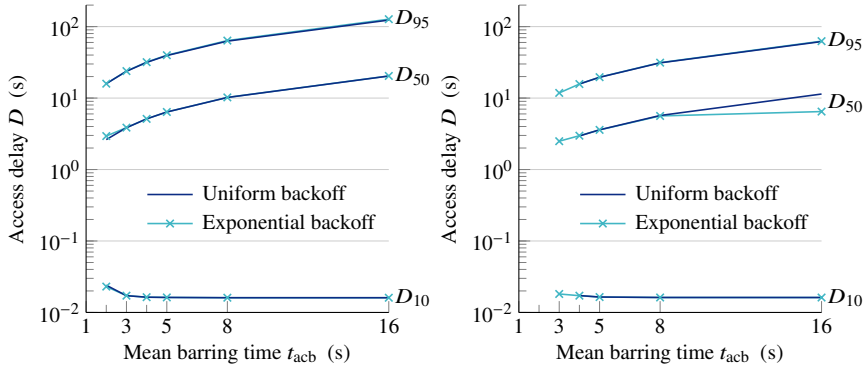


Figure 2.12: Percentiles of access delay D of M2M UEs under the ACB scheme for the combinations of t_{acb} with (a) $p_{acb} = 0.3$ and (b) $p_{acb} = 0.5$ that lead to $P_s \geq 0.95$.

of the values of D_{10} with the minimum time needed to complete the RAP, which is 15 ms (see Table 2.3 on page 26). On the other hand, $D_{10} \approx 15$ ms, $D_{50} \approx 20$ ms, and $D_{95} \approx 60$ ms were obtained for H2H UEs and for any combination of the barring parameters. These values are less than 10 percent away from those obtained when a single UE access occurs. This demonstrates that high priority traffic is not affected by mMTC applications given that the ACB scheme successfully prevents congestion.

On the other hand, a clear tradeoff can be identified by comparing the results presented in Fig. 2.11 with those in Fig. 2.12: combinations of the barring parameters that lead to a low $\mathbb{E}[K]$ result in a long access delay. Naturally, the reason for this is that combinations of parameters that lead to a sharp drop in the number of contending UEs per RAO also decrease $\mathbb{E}[K]$ significantly. However, a long access delay is needed to sharply decrease the number of contending UEs per RAO. Building on this, barring parameters must be carefully selected to meet with the performance requirements of the target application. In the following, we assume D is the second most important KPI and seek to find the optimal configuration of the ACB scheme, defined as the combination of barring parameters that minimizes D_{95} while achieving $P_s \geq 0.95$.

Fig. 2.13 shows the cumulative distribution function (CDF) of D for two of the

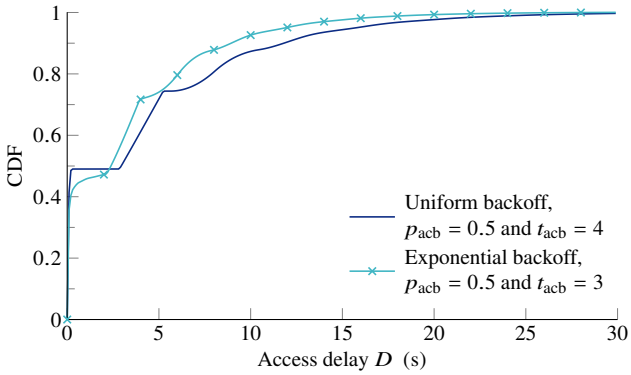


Figure 2.13: CDF of access delay for the combinations of barring parameters and backoff implementations that lead to the shortest D_{95} given $P_s \geq 0.95$.

combinations of barring parameters, among those that have been considered for the results presented in Fig. 2.11 and Fig. 2.11, that lead to $P_s \geq 0.95$ with the minimum D_{95}. Please recall that so far the possible values of p_{acb} and t_{acb} have been restricted to those available for selection in the SIB 2. Results shown in Fig. 2.13 include one uniform backoff and one exponential backoff implementation. As observed previously, a shorter D is achieved with the exponential backoff.

Our performance analysis of the ACB scheme by simulation concludes with the identification of the optimal configuration for a vast collection of values for p_{acb} and t_{acb} that are not restricted to those included in the SIB 2. For this, we first identify the optimal value of t_{acb}, denoted as t^*_{acb} for $p_{\text{acb}} \in \{0.01, 0.02, \ldots, 0.99\}$. That is, the value of t_{acb} that minimizes D_{95} for a given p_{acb}; we have observed that it also corresponds to the minimum t_{acb} that leads to $P_s \geq 0.95$. Hence, t^*_{acb} can be defined as follows.

$$t^*_{\text{acb}} = \min \{t_{\text{acb}} \mid P_s\,(p_{\text{acb}}, t_{\text{acb}}) \geq 0.95\} \tag{2.14}$$

As described above, $P_s \geq 0.95$ cannot be achieved when $p_{\text{acb}} \gg 0.5$. Specifically, the greatest value of p_{acb} for which $P_s \geq 0.95$ was obtained was 0.56 with the uniform backoff and 0.62 with the exponential backoff. Fig. 2.14 shows t^*_{acb} as a function of p_{acb}. Then, Fig 2.15 shows $\mathbb{E}[K]$ and D^*_{95}, where the tradeoff between these two

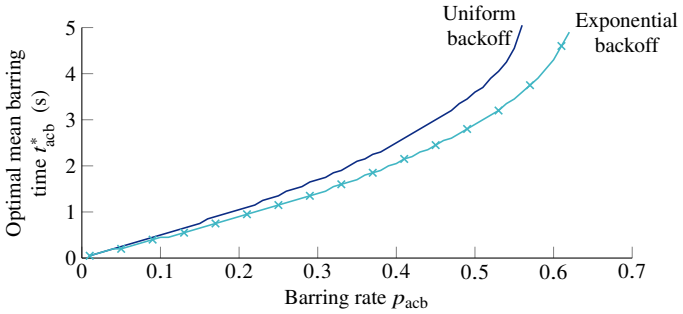


Figure 2.14: Optimal mean barring time t^*_{acb} given p_{acb}.

parameters is evident. Also, form Fig. 2.15 the optimal value of p_{acb}, defined as

$$p^*_{acb} = \arg\min D_{95}\left(p_{acb}, t^*_{acb}\right) \tag{2.15}$$

can be obtained. For the uniform backoff $p^*_{acb} = 0.31$ and $t^*_{acb} = 1.75$; for the exponential backoff $p^*_{acb} = 0.35$ and $t^*_{acb} = 1.7$. The resulting D_{95} is 13.554 and 11.239 s, respectively.

2.5 Conclusions

This chapter presented the description and the theoretical capacity of the RAP. Furthermore, the ACB scheme was described and its benefits were investigated. Results were obtained mainly by simulations and demonstrate that the resources available at the PRACH and at the PDCCH are not sufficient to handle the large number of synchronized UE arrivals that occur in mMTC. We observed that manipulating the configuration parameters such as the maximum number of preamble transmissions per UE may maximize the utilization of available resources, but, since these are not sufficient, the achieved performance was not acceptable with any of the studied parameter combinations. Furthermore, the implementation of a different backoff favorably impacts performance but is also insufficient to relieve congestion and to achieve the target success probability of 0.95.

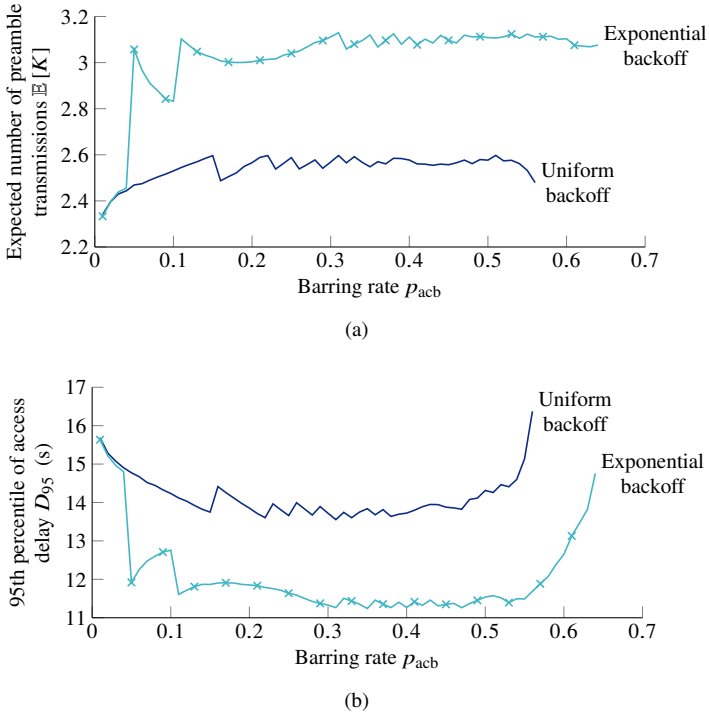


Figure 2.15: Achieved (a) $\mathbb{E}[K]$ and (b) D^*_{95} given t^*_{acb}.

On the other hand, the ACB scheme is a promising solution to congestion under mMTC applications as even a simple implementation, in which barring parameters remain constant throughout the distribution period, can lead to the desired success probability. In addition, a careful selection of barring parameters may lead to a balance between access delay and preamble transmissions that satisfies the application requirements. However, it is important to emphasize that the achieved access delay and, more specifically, the minimum achievable 95th percentile of access delay is longer than 10 s. As a consequence, the ACB scheme is an appealing solution to congestion, but is only suitable for delay tolerant applications.

Chapter 3

Analytical modeling of RA in cellular networks

3.1 Introduction

he capacity of the RA procedure (RAP) in cellular networks and the benefits of the access class barring (ACB) scheme to relieve congestion under massive machine-type communication (mMTC) scenarios were investigated in Chapter 2. In particular, we focused on the scenario that results in the highest access intensity according to studies performed by the 3rd Generation Partnership Project (3GPP) on dense urban environments: the traffic model (TM) 2 [1]. Results presented in Chapter 2 are in line with those obtained in the literature [1, 34, 40] and demonstrate that congestion is likely to occur under highly synchronized mMTC scenarios. For instance, only $P_s = 0.313$ of the user equipments (UEs) successfully complete the RAP when $n = 30\,000$ UEs access the evolved NodeB (eNB) according to the TM 2. Hence, the RAP is not efficient for mMTC applications.

The main reasons for achieving such a poor performance are the complexity of the RAP and the limitations of its uplink and downlink physical channels. Nevertheless, we also observed in Chapter 2 that the performance can be greatly enhanced by implementing and correctly configuring the ACB scheme. The ACB is an access

control scheme that redistributes the UE arrivals through time, so that the capacity of the RAP is not exceeded. That is, the ACB scheme is effective given adequate configuration parameters are selected.

Up to this point, most of the results on the performance of the random access (RA) (including the ACB scheme) that have been presented have been obtained by means of simulations, which have a high computational cost and, more importantly, are not easily reproducible. Instead, an analytical model would provide with much more computational efficiency, reproducibility, and with a much more in depth understanding of its behavior; hence, having an analytic model of the RAP at our disposal was of prime importance for our research.

Our search for an analytical model in the literature revealed one of the first efforts to model the RAP was presented by Zhou *et al.* in 2008 [83], but only the first step, preamble transmission, was considered. In fact, we found that there were just a few analytic models of the complete RAP and, as it will be seen in Section 3.4, their accuracy suffers when compared to simulations [20, 30, 109].

As a starting point, we decided to use the model provided by Wei *et al.* [109]; hereafter we refer to this as the reference model (RM). To the best of our knowledge, this was the most thorough analytic model for the performance evaluation of the RAP, and was later extended to incorporate an access control scheme called extended access barring [30]. Therefore, the basic model of the RAP presented by Cheng *et al.* is the same as the RM. Furthermore, the model presented by Arouk *et al.* [20] is also of similar nature to that of the RM, but only the average delay is calculated in the former while in the latter, the probability mass function (pmf) is calculated.

Nevertheless, we observed the RM is inaccurate, especially when the number of successful accesses per random access opportunity (RAO) approaches the RAP capacity C. That is, when most resources are utilized. This is an enormous downside because, as we observed in Chapter 2 and will observe in Chapter 4, the point of optimal operation of the RAP under massive mMTC scenarios is when most resources are being utilized. Therefore, the main objective of access control schemes is to maintain the system at this optimal point. Clearly, we cannot rely on an analytic model that fails to deliver the required accuracy when it is most needed.

Building on this, we decided to study the components that contribute to reduce the accuracy of the RM. We reached the same conclusion as Arouk *et al.* [20]: the accuracy of the RM is mainly reduced by exclusively using the expected values of several random variables (RVs). Instead, using the whole pmfs seems much more appropriate. Then, a new model of the RAP with this distinctive characteristic was developed.

As mentioned above, an access control is needed to support mMTC in cellular networks. Therefore, the ACB scheme was also envisioned to be incorporated to our model. As described in Chapter 2, oftentimes the behavior of the ACB scheme is misinterpreted. For example Lin *et al.* [71] assume the time UEs must wait after each failed barring check (i.e., the barring time) is fixed, whereas the technical specifications clearly state that it is calculated randomly [10]. On the other hand, our preliminary work [68] was one of the first studies in the literature to evaluate the benefits of the ACB scheme with the behavior described by the 3GPP [10, Section 5.3.3.11]. The information collected during this process with respect to the behavior of the ACB scheme led to the development of an analytic model of the ACB scheme that, in combination with the analytic model of the RAP, has allowed us to thoroughly evaluate the complete RA.

In this chapter, we present our novel analytic model for the performance evaluation of the RA in cellular networks. By means of this model, the following key performance indicators (KPIs) (selected from the ones suggested by the 3GPP [1]) can be accurately calculated:

1. Success probability, defined as the probability to successfully complete the RAP within the maximum number of preamble transmissions.
2. Collision probability, defined as the ratio between the total number of preambles transmitted simultaneously by multiple UEs and the total number of available preambles in the period in which accesses occur.
3. Probability distribution of the number of preamble transmissions performed by the UEs that successfully complete the RAP.
4. Probability distribution of the access delay.

We assess the accuracy of our model by comparing the results obtained with both, our model and the RM with those obtained by simulation. As it will be observed throughout Section 3.4, our model surpasses the accuracy of the RM in the vast majority of KPIs and network configurations, and has allowed us to reach important conclusions regarding the behavior of the RAP. In addition, results can be obtained with our model within a few tens of seconds for the selected scenario. These results in turn have served us as a base to develop the adaptive access class barring configuration (ACBC) scheme presented in Chapter 4.

An important aspect regarding the behavior of the RAP is the lack of consensus in the literature on the outcomes and assumptions regarding preamble collisions during preamble transmission. That is, there are several causes for the eNB to decode or not decode preambles transmitted by multiple UEs simultaneously; hereafter we refer to these as collided preambles. These possible causes and their implications are also studied in this chapter.

The rest of the chapter is organized as follows. In the following section, we summarize the RAP as defined by the 3GPP. Please refer to Section 2 for an in-depth description of the RAP. Then, we describe the two different outcomes that may occur when multiple UEs transmit the same preamble simultaneously, along with the most common assumptions in the literature regarding these outcomes. Then, our analytical model is presented in Section 3.3. Next, Section 3.4 presents relevant results on the accuracy of our model and on the performance of the RA under different network configurations. These results include the performance of the RAP under the two different assumptions regarding the outcome of preamble collisions. Finally, we present our conclusions in Section 3.5.

3.2 RA in cellular networks: possible outcomes and common assumptions

This section provides a brief description of the ACB scheme and of the RAP defined by the 3GPP [5, 10] for the initial access in cellular networks, along with a detailed

description on the possible outcomes and common assumptions regarding preamble collisions.

Upon arrival, UEs obtain the configuration of the network through System Information Blocks (SIBs). These include, among others, the number of available orthogonal preambles available, these are Zadoff-Chu sequences in LTE Advanced (LTE-A) and single tone frequencies in narrowband Internet of Things (NB-IoT), the period of time-frequency resources in which preamble transmissions are allowed (known as RAOs), and barring parameters. With this information and before initiating the RAP, the UEs are subject to the ACB scheme. That is, before initiating the transmission of the first preamble, UEs must succeed in a barring check. This occurs with probability equal to the barring rate; otherwise, the UE must wait for a period calculated randomly.

UEs that succeed in a barring check are allowed to initiate the RAP, which comprises a four-message handshake. The first message is preamble transmission (*Msg1*), where preambles are selected randomly and transmitted in RAOs. Then, the eNB may send up to one uplink grant in response to each preamble decoded successfully (*Msg2*). The eNB can only respond to preambles transmitted in a RAO during the next RA response (RAR) window, whose duration is oftentimes equal to the period of RAOs. UEs that receive an uplink grant send a connection request (*Msg3*) through dedicated resources and, finally, the eNB responds with the contention resolution message (*Msg4*). *Msg3* and *Msg4* are protected with hybrid ARQ (HARQ) mechanisms. Therefore, several retransmissions of these messages are attempted until *Msg4* is received or until the maximum number of attempts is reached. Whenever the latter occurs or if preamble transmission fails and if the maximum number of preamble transmissions has not been exceeded, the UE waits for a random backoff time, increases the transmission power, and then transmits a newly selected preamble at the next RAO.

Preambles are orthogonal sequences. Therefore, it is clear that multiple UEs can access the eNB at the same RAO, using different preambles. Also, it is clear that preambles transmitted by exactly one UE, hereafter denoted as successful preambles, and transmitted with sufficient power are decoded by the eNB. On the other hand, the reasons for a RA failure are manifold and give rise to two different outcomes when multiple UEs transmit the same preamble simultaneously; hereafter we refer to these as

collided preambles. These two outcomes and the circumstances involved are described in the following.

- **The eNB does not decode the transmitted preamble.** This may occur if the eNB determines a preamble was transmitted by multiple UEs. This case can be identified, for example, based on the received signal power and the time shift between the multiple received copies of the preamble. A different cause for this outcome is that all preamble transmissions are lost due to wireless channel errors, and can occur, for example, if the interference caused by the multiple preamble transmissions is exceedingly high. Regardless of the cause, the UEs will not receive an uplink grant by the end of the next RAR window; it is at this point in time that the implicated UEs will detect the collision.

- **The eNB correctly decodes the transmitted preamble.** This outcome may occur if the received power from one of the preamble transmissions is significantly higher than that of the other simultaneous transmissions of the same preamble, commonly known as the capture effect. Other possible cause for this outcome is that all but one of these preamble transmissions are lost due to wireless channel errors. In both cases, the multiple UEs that transmitted the preamble will receive the exact same uplink grant and continue with the RAP by sending *Msg3*. Then, the eNB will receive multiple *Msg3*s with different data at the exact same time-frequency resources and will not transmit *Msg4* in response. As a consequence, the preamble collision will be detected until the maximum number of *Msg3* transmission attempts is reached without success.

Naturally, in a real life scenario either of these two outcomes may occur with a given probability. However, this probability is not known and highly depends on the wireless environment of the cell of interest. Therefore, the 3GPP recommends to assume collided preambles are never decoded by the eNB [1]. This assumption has been adopted in most of the literature [20, 30, 71, 94, 109]. Nevertheless, some studies assume the opposite. That is, collided preambles are always decoded by the eNB [33–35, 79]. Throughout this thesis we assume collided preambles are never decoded by the eNB as suggested by the 3GPP, but also investigate the impact on performance of

assuming the opposite. Specifically, the analytical model presented in the following section was designed under the assumption that collided preambles are never decoded by the eNB. Therefore, most of the results presented in Section 3.4 were obtained under this assumption. results on the performance of the RAP given the opposite assumption are presented in Section 3.4.

3.3 Analytical model of the RA in cellular networks

This section presents the analytical model for the performance evaluation of the RA in cellular networks, which includes the RAP and the ACB scheme. For illustration purposes, we assume the most typical configuration of the physical RACH (PRACH) and physical downlink control channel (PDCCH) in LTE-A, along with its timing parameters. That is, the default value of the period of RAOs is $t_{\text{rao}} = 5$ ms unless otherwise stated. Table 3.1 presents other important parameters used throughout this chapter. Needless to say, our model can be easily adapted to other configurations and timing parameters. For example, to typical values in NB-IoT.

3.3.1 Modeling the UE arrivals

Let RV A define the number of RAOs elapsed between the beginning of the distribution period $i = 0$ and the arrival of a specific UE. That is, the RAO in which the UE schedules the beginning of its RAP; the pmf of A for each UE is given by the selected traffic model [1].

Under the TM 2, UEs arrivals follow a Beta $(3, 4)$ distribution over $t_{\text{dist}} = 10$ s [1].

Definition 3.3.1. *The probability density function (pdf) of a continuous RV* $T \sim \text{Beta}(\alpha, \beta)$ *whose support is* $t \in [0, 1]$*, is defined as follows.*

$$f(t; \alpha, \beta) = \frac{t^{\alpha-1}(1-t)^{\beta-1}}{\displaystyle\int_0^1 v^{\alpha-1}(1-v)^{\beta-1}\,\mathrm{d}v} = \frac{t^{\alpha-1}(1-t)^{\beta-1}}{\mathrm{B}(\alpha, \beta)} \tag{3.1}$$

Table 3.1: Parameters for the selected PRACH and PDCCH configuration and TM 2.

Parameter	**Setting**
Number of machine-to-machine (M2M) UEs	$n = 30\,000$
Distribution period	$t_{\text{dist}} = 10$ s
Distribution of UE arrivals	Beta $(3, 4)$
Subframe length	$t_{\text{s}} = 1$ ms
Subframe frequency	$f_{\text{s}} = 1$ subframes per ms
Period of RAOs	$t_{\text{rao}} \in \{2, 5, 10\}$ ms
RAR window size	$t_{\text{rar}} = t_{\text{rao}}$ ms
Available preambles	$r = 54$
Maximum frequency of uplink grants	$f_g = 3$ per ms
Available uplink grants per RAR window	$g = f_g t_{\text{rar}}$
Maximum number of preamble transmissions	$k_{\max} = 10$
Backoff Indicator	$b_{\max} = 20$ ms
ACB barring rate	$p_{\text{acb}} = 0.5$
ACB barring time	$t_{\text{acb}} = 4$ s
Preamble processing delay	$d_{\text{p}} = 2$ ms
Uplink grant processing delay	$d_{\text{ug}} = 5$ ms
Msg3 processing delay	$d_{\text{cr}} = 4$ ms
Msg3 round-trip time (RTT)	$d_{\text{m3}} = 8$ ms
Msg4 RTT	$d_{\text{m4}} = 5$ ms
Maximum number *Msg3* and *Msg4* transmissions	$h_{\max} = 5$
Error probability for the kth preamble transmission	$\Pr[\mathcal{E}_k] = 1/e^k$
Error probability for *Msg3* and *Msg4* transmissions	$\Pr[\mathcal{E}_h] = 0.1$

where $\mathrm{B}(\cdot)$ *is the Beta function, defined for any* $\alpha, \beta \in \mathbb{N}$ *as*

$$\mathrm{B}(\alpha, \beta) = \frac{(\alpha-1)!\,(\beta-1)!}{(\alpha+\beta-1)!} \tag{3.2}$$

Next, let X_i be the RV that defines the the number of UE arrivals at the ith RAO; hence $\{X_i\}_{i\in\mathbb{N}}$ is a stochastic process. In most of the literature, the expected value of X_i is obtained as follows [1, 30].

$$\mathbb{E}[X_i] = n \int_{t_i}^{t_i+1} f\left(\frac{t}{t_{\mathrm{dist}}}; \alpha, \beta\right) \mathrm{d}t \tag{3.3}$$

where $t_i = it_{\mathrm{rao}}$ and $f(t/t_{\mathrm{dist}}; \alpha, \beta)$ is the pdf of a transformed version of RV T with support $t \in [0, t_{\mathrm{dist}}]$. According to the 3GPP [1], this pdf is given as

$$f\left(\frac{t}{t_{\mathrm{dist}}}; \alpha, \beta\right) = \frac{t^{\alpha-1}\,(t_{\mathrm{dist}}-t)^{\beta-1}}{t_{\mathrm{dist}}^{\alpha-\beta-1}\,\mathrm{B}(\alpha, \beta)}. \tag{3.4}$$

In the most widely studied mMTC scenario, UE arrivals follow the TM 2, where $\alpha = 3$, $\beta = 4$, and $t_{\mathrm{dist}} = 10$ s. But obtaining $\mathbb{E}[X_i]$ directly from the previous formulations is cumbersome. Instead, the following approach to calculate $\mathbb{E}[X_i]$ directly from a discrete Beta $(3, 4)$ distribution, is proposed.

Lemma 3.1. *Let* $i_{\mathrm{dist}} = t_{\mathrm{dist}}/t_{\mathrm{rao}}$ *be the last RAO within the distribution period. As such, the support of A is* $\{i \in \mathbb{N} \mid i \le i_{\mathrm{dist}}\}$ *and its pmf under the TM 2 can be given as*

$$p_A(i) = \frac{60i^2\,(i_{\mathrm{dist}}-i)^3}{i_{\mathrm{dist}}^6 - i_{\mathrm{dist}}^2}. \tag{3.5}$$

From there it follows immediately that

$$\mathbb{E}[X_i] = np_A(i) = \frac{60ni^2\,(i_{\mathrm{dist}}-i)^3}{i_{\mathrm{dist}}^6 - i_{\mathrm{dist}}^2} \qquad \text{for } i \in \{0, 1, \ldots, i_{\mathrm{dist}}\}. \tag{3.6}$$

The proof of Lemma 3.1 is in Appendix B.2.

The three chapters of this thesis dedicated to mMTC in cellular networks focus on the performance evaluation of the RAP under typical configurations of the PRACH. The

most typical value of the period of RAOs is $t_{\text{rao}} = 5$ ms; hence, this is the value selected from Chapter 2 to Chapter 4. In addition, the impact of selecting $t_{\text{rao}} \in \{2, 10\}$ ms is also studied in this chapter. For $t_{\text{rao}} \in \{2, 5, 10\}$ $i_{\text{dist}} = \{5000, 2000, 1000\}$, which takes us to the following proposition used in our previous work [60, 69].

Proposition 3.1. *It is immediate to see that* $a/\left(i_{\text{dist}}^6 - i_{\text{dist}}^2\right) \approx a/i_{\text{dist}}^6$ *for sufficiently large* i_{dist}. *Building on this,* (3.6) *can be expressed in an even simpler form that matches its continuous version as defined by* (3.4) *as*

$$p_A(i) \approx \frac{60\, i^2\, (i_{\text{dist}} - i)^3}{i_{\text{dist}}^6}. \tag{3.7}$$

Therefore the expected value of X_i *can be calculated as follows.*

$$\mathbb{E}[X_i] \approx \frac{60\, n\, i^2\, (i_{\text{dist}} - i)^3}{i_{\text{dist}}^6} \qquad \text{for } i \in \{0, 1, \ldots, i_{\text{dist}}\}. \tag{3.8}$$

Now the critical values of $\mathbb{E}[X_i]$ presented in Section 2.4 can be confirmed; these values were previously obtained by simulation to estimate the minimum number of available preambles r that may lead to $P_s \geq 0.95$. To confirm these values, let i^* be the value of i that maximizes $\mathbb{E}[X_i]$; it can be easily obtained by the first derivative test as follows.

$$\frac{\partial\, \mathbb{E}[X_i]}{\partial i} = \frac{60\, n \left(2i\, (i_{\text{dist}} - i)^3 - 3i^2\, (i_{\text{dist}} - i)^2\right)}{i_{\text{dist}}^6} = 0 \tag{3.9}$$

which gives $i^* = (2i_{\text{dist}})/5$. Naturally, this result matches the mode of a Beta $(3, 4)$ distribution with support $t \in [0, 1]$ but shifted to the right by a factor of i_{dist}. The maximum expected number of UE arrivals is $\mathbb{E}[X_{i^*}] = 31.104$, the exact same value obtained by simulation.

3.3.2 Modeling the ACB scheme

Once we have obtained the distribution of UE arrivals we proceed to present the model of the ACB scheme. In this chapter, similarly as in Chapter 2, we assume every UE is subject to the ACB scheme with fixed parameters. That is, the barring rate $p_{\text{acb}}(j)$ and

the barring time $t_{\text{acb}}(j)$ remain constant for all $j \in \mathbb{N}$ SIB 2 transmissions. Hence, to simplify notation we simply denote these as p_{acb} and t_{acb}, respectively.

Next, let W be the RV that defines the number of RAOs that the first preamble transmission of a UE is delayed due to the ACB scheme. That is, the number of RAOs that a UE has to wait to begin the RAP due to failed barring checks. Hence, the sample space of W is $i \in \mathbb{N}$. Also, let Y be the RV with sample space $y \in \mathbb{Z}_+$ that represents the number of barring checks performed by a UE. It is clear that the preamble is transmitted immediately if the UE succeeds in its first barring check. That is, $\Pr[W = 0 \mid Y = 1] = 1$, which occurs with probability p_{acb}. At this point, we define the function

$$\delta(i) \equiv \begin{cases} 1, & i = 0 \\ 0, & \text{otherwise} \end{cases} \tag{3.10}$$

which allows us to easily define the pmf of $W \mid Y = 1$ as

$$p_{W|Y}(i \mid 1) = \delta(i). \tag{3.11}$$

From there, it is clear that the pmf of $W \mid Y = 2$ is positive between $i_{W,\min} = \lceil 0.7\, t_{\text{acb}}/t_{\text{rao}} \rceil$ and $i_{W,\max} = \lceil 1.3\, t_{\text{acb}}/t_{\text{rao}} \rceil$. Its pmf is given as

$$p_{W|Y}(i \mid 2) = \frac{1}{0.6\, t_{\text{acb}}} \begin{cases} i\, t_{\text{rao}} - 0.7\, t_{\text{acb}}, & i = i_{W,\min} \\ t_{\text{rao}}, & i_{W,\min} < i < i_{W,\max} \\ 1.3\, t_{\text{acb}} - (i-1)\, t_{\text{rao}}, & i = i_{W,\max}, \end{cases} \tag{3.12}$$

then, the pmf of $W \mid Y$ can be calculated recursively as

$$p_{W|Y}(i \mid y) = \sum_{v=i_{W,\min}}^{i_{W,\max}} p_{W|Y}(v \mid 2)\; p_{W|Y}(i - v \mid y - 1), \qquad y = 3, 4, \ldots \tag{3.13}$$

Naturally, each barring check is a single Bernoulli trial. Hence, Y is a geometric RV whose pmf is given as

$$p_Y(y) = p_{\text{acb}}(1 - p_{\text{acb}})^{y-1}, \qquad \text{for } y = 1, 2, \ldots \tag{3.14}$$

Now we are able to calculate the pmf of W as

$$p_W(i) = \sum_{y=1}^{\infty} p_{W|Y}(i \mid y)\, p_Y(y), \qquad \text{for } i = 0, 1, 2, \ldots \tag{3.15}$$

Please observe that the support of both W and Y is infinite, which impedes their inclusion in out model. To circumvent this problem, we define $y_{\max}$ as the maximum number of barring checks performed by the UEs. Therefore, the pmf of Y is truncated at $y_{\max}$. This implies that, in our model, UEs that fail the first $y_{\max}$ barring checks desist and consider the RA as terminated without success (i.e., failed). The probability that a given UE fails the first $y_{\max}$ barring checks and terminates the RA is simply

$$p_{\mathcal{E}_{\text{acb}}} = (1 - p_{\text{acb}})^{y_{\max}}. \tag{3.16}$$

As such, $y_{\max}$ can be calculated for a target $p_{\mathcal{E}_{\text{acb}}}$ from (3.41) as

$$y_{\max} = \left\lceil \frac{\log p_{\mathcal{E}_{\text{acb}}}}{\log(1 - p_{\text{acb}})} \right\rceil, \tag{3.17}$$

where $p_{\mathcal{E}_{\text{acb}}}$ is selected empirically. This allows us to approximate $p_W(i)$ by truncating (3.15) as

$$p_{\hat{W}}(i) = p_{W|Y \le y_{\max}}(i) = \frac{1}{1 - p_{\mathcal{E}_{\text{acb}}}} \sum_{y=1}^{y_{\max}} p_{W|Y}(i \mid y)\, p_Y(y). \tag{3.18}$$

Please observe that $p_{\hat{W}}(i)$ is indeed a probability distribution and that $p_W(i) \to p_{\hat{W}}(i)$ when $p_{\mathcal{E}_{\text{acb}}} \to 0$. Throughout this chapter we select $p_{\mathcal{E}_{\text{acb}}} = 10^{-5}$ as the target probability that a UE terminates the RA during a barring check and denote $p_{\hat{W}}(i)$ simply as $p_W(i)$. From there, let D_{ACB} be the RV that defines the delay due to the ACB scheme in seconds. It can be easily calculated as the following function of W.

$$\Pr[D_{\text{ACB}} = i\, t_{\text{rao}}] = \Pr[W = i]. \tag{3.19}$$

Fig. 3.1 shows $F_W(i)$, the cumulative distribution function (CDF) of W for the target $p_{\mathcal{E}_{\text{acb}}} = 10^{-5}$ and two combinations of barring parameters. The first one is $p_{\text{acb}} = 0.5$, $t_{\text{acb}} = 4$ s and corresponds to the optimal configuration of the ACB with fixed values available for selection in the SIB 2. The second one is $p_{\text{acb}} = 0.31$,

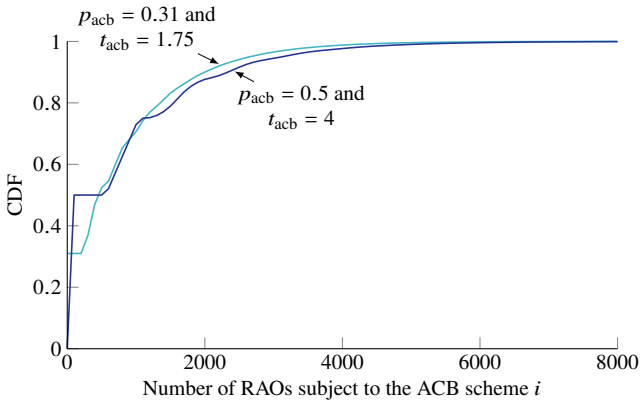


Figure 3.1: CDF of the barring time W in RAOs for $p_{\text{acb}} = 0.5$, $t_{\text{acb}} = 4$, and $\Pr[\mathcal{E}_{\text{acb}}] = 10^{-5}$.

$t_{\text{acb}} = 1.75$ s and corresponds to the optimal configuration found in Chapter 2 for an expanded collection of values of p_{acb} and t_{acb}.

Now that we have obtained the pmf of the UE arrivals $p_A(i)$ and of the barring time $p_W(i)$ we proceed to calculate the number of UEs that begin the RAP at each RAO.

Please recall that k is the number of preamble transmissions performed by a UE and that $k_{\max}$ is the maximum number of transmissions allowed and is transmitted by the eNB in the SIB 2; hence, $k \in \{1, k_{\max}\}$. Next, let $N_i(k)$ be the RV that defines the number of UEs that perform their kth preamble transmission at the ith RAO. The expected number of UEs that perform their first preamble transmission (i.e., begin the RAP) at the ith RAO is given as

$$\mathbb{E}[N_i(1)] = n \Pr[A + W = i] = n \sum_{v=0}^{i} p_A(v)\ p_W(i - v)\,. \tag{3.20}$$

Please observe that $\mathbb{E}[N_i(1)] = \mathbb{E}[X_i]$ if $\Pr[W = 0] = 1$, which occurs when the ACB scheme is disabled (i.e., when $p_{\text{acb}} = 1$). The expected number of UEs that are about to perform their kth preamble transmission will later be obtained recursively by means of (3.37) on page 61.

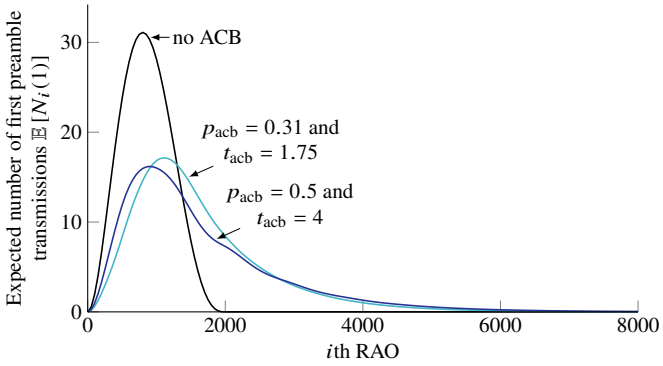


Figure 3.2: Expected number of first preamble transmissions per RAO $\mathbb{E}[N_i(1)]$ under the TM 2 with $n = 30\,000$ for three cases: 1) disabled ACB scheme; 2) ACB with $p_{acb} = 0.5$ and $t_{acb} = 4$; and 3) ACB with $p_{acb} = 0.31$ and $t_{acb} = 1.75$.

Fig. 3.1 shows $\mathbb{E}[N_i(1)]$ for three configurations. In the first one, no ACB scheme is implemented. This same behavior can be achieved by selecting $p_{acb} = 1$. In the second one $p_{acb} = 0.5$ and $t_{acb} = 4$. In the third one $p_{acb} = 0.31$, $t_{acb} = 1.75$. It can be sen that the latter ACB configuration leads to a higher $\mathbb{E}[N_i(1)]$ than $p_{acb} = 0.5$ and $t_{acb} = 4$, which, as observed in Chapter 2, also leads to a shorter access delay while achieving $P_s \geq 0.95$. Throughout the rest of this chapter we focus on evaluating the performance of the RA with $p_{acb} = 0.5$ and $t_{acb} = 4$, as it is the optimal configuration with values provided in the SIB 2. The performance achieved with the combination $p_{acb} = 0.31$ and $t_{acb} = 1.75$ will be further investigated in Chapter 4. We now proceed to describe our analytical model of the RAP.

3.3.3 Modeling the RAP

The analytical model that is presented in the following has been developed under the assumption that the eNB can only decode preambles transmitted by exactly one UEs at the same RAO. This goes in line with the 3GPP recommendations for the performance evaluation of the RA procedure [1], with most of the literature [20, 30, 71, 94, 116].

Let N_i be the RV that defines the number of contending UEs at the ith RAO. The

sample space of N_i is $n(i) \in \mathbb{N}$ and its expected value can be obtained as

$$\mathbb{E}[N_i] = \sum_{k=1}^{k_{\max}} \mathbb{E}[N_i(k)]\,. \tag{3.21}$$

As a starting point, we obtain the pmfs of preambles transmitted by exactly one (successful transmissions) and by multiple UEs (collisions) given $n(i) \in \mathbb{N}$. Then, we derive these same pmfs for any $\mathbb{E}[N_i] \in \mathbb{R}_{\geq 0}$

The process of preamble selection and transmission can be modeled as a bins and balls problem, in which a given number of balls is placed randomly in one out of the available bins [109]. Please recall that r is the number of available preambles and let S_i and C_i be the RVs that define the number of successful preambles (i.e. transmitted by exactly one UE) and collided preambles (i.e., transmitted by multiple UEs) at the ith RAO, respectively. Therefore $\{S_i\}$ and $\{C_i\}$ are stochastic processes.

Back to the bins and balls problem, $\mathbb{E}[N_i]$ represents the number of balls and r the number of bins, whereas S_i and C_i represent the number of bins with exactly one ball and the number of bins with more than one ball at the ith realization of the experiment, respectively. The sample space of S_i is the number of successes $\mathcal{S} = \{s \in \mathbb{N} \mid 0 \leq s \leq s_{\max}\}$, where $s_{\max} = \min\{r, \mathbb{E}[N_i]\}$. On the other hand, the sample space of C_i is the number of collided preambles $C = \{c \in \mathbb{N} \mid 0 \leq c \leq c_{\max}\}$, where $c_{\max} = \min\{r, \mathbb{E}[N_i]/2\}$.

Proposition 3.2. *To solve this problem efficiently, we first define the auxiliary RVs S and C; these are analogous to S_i and C_i, but defined for an arbitrary experiment and for $m \in \mathbb{N}$. Then, we calculate the joint probability distribution of S AND C for a given m recursively as follows.*

$$\begin{aligned} p_{S,C}(s,c;m) = &\left(\frac{r-s+1-c}{r}\right) p_{S,C}(s-1,c;m-1) + \frac{c}{r}\, p_{S,C}(s,c;m-1) \\ &+ \frac{s+1}{r}\, p_{S,C}(s+1,c-1;m-1) \qquad \forall s \in \mathcal{S} \text{ and } c \in C \end{aligned} \tag{3.22}$$

given the initial condition $p_{S,C}(0,0;0) = 1$.

That is, we derive the probability of having s successful and c collided preambles for a given discrete value m from the case in which $m-1$ UEs have already selected their preamble. For this, three possibilities exist:

- $s-1$ preambles are selected by exactly one and c preambles are selected by multiple UEs ; then a new UE selects any of the $r-(s-1)-c$ preambles that have not been selected by other UEs.
- s preambles are selected by exactly one and c preambles are selected by multiple UEs ; then a new UE selects one of the c preambles.
- $s+1$ preambles are selected by exactly one and $c-1$ preambles are selected by multiple UEs ; then a new UE selects one of the $s+1$ preambles.

The marginal pmfs of S and C are easily calculated as

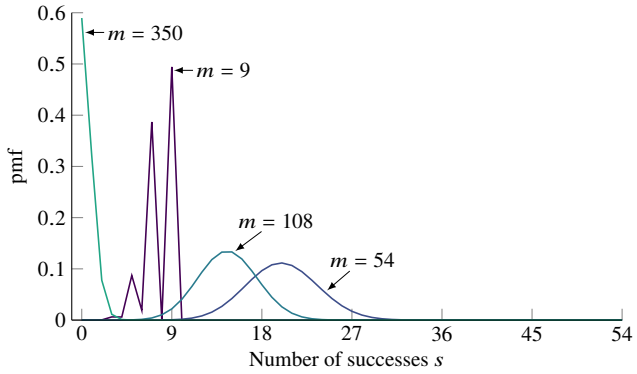
$$p_S(s;m) = \sum_{c=0}^{c_{\max}} p_{S,C}(s,c;m) \tag{3.23}$$

$$p_C(c;m) = \sum_{s=0}^{s_{\max}} p_{S,C}(s,c;m)\,. \tag{3.24}$$

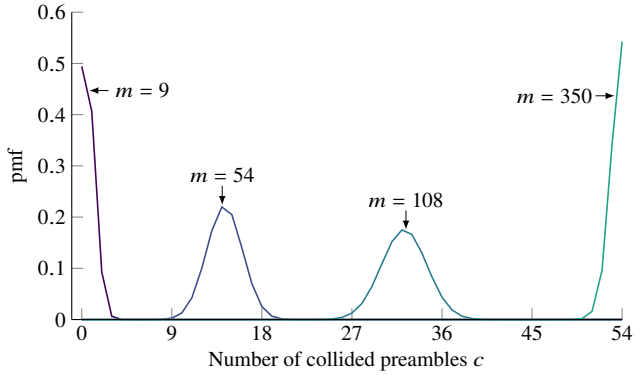
These can be calculated once for a sufficiently large m and stored in a two-dimensional matrix for further use. For example, we have observed that under the TM 2 with no ACB scheme and $t_{\text{rao}} = 5$ ms, the maximum expected number of contending UEs is around 300. Hence, calculating the pmf of S AND C for all $m = 1, 2, \ldots, 350$ is sufficient.

In recent studies [96], we have compared the computational complexity of (3.22) with that of other formulations presented in the literature [20, 35, 109]. Results show (3.22) is up to 6000 and at least 70 times faster than the other formulations found in the literature for $r = 54$ and $m = 400$.

Fig. 3.3 shows the pmfs of S and C given $r = 54$ for characteristic values of m. Clearly, the pmf of S for low values of m is highly variable, but tends to a normal distribution for $m \approx r$. On the other hand, $\Pr[S=0] \to 1$ and $\Pr[C=r] \to 1$ as $m \to \infty$.



(a)



(b)

Figure 3.3: Pmf of the number of (a) successful S and (b) collided C preambles for $r = 54$ and $m \in \{9, 54, 108, 350\}$ contending UEs.

Now we are able to derive the pmf of S_i from the marginal pmf of S by means of the linear interpolation

$$p_{S_i}(s) = p_S(s; \lceil \mathbb{E}[N_i] \rceil)\ (\mathbb{E}[N_i] - \lfloor \mathbb{E}[N_i] \rfloor) + p_S(s; \lfloor \mathbb{E}[N_i] \rfloor)\ (1 - \mathbb{E}[N_i] + \lfloor \mathbb{E}[N_i] \rfloor)\,. \tag{3.25}$$

The pmf of C_i can be derived from the marginal pmfs of C analogously.

As described in the previous section, preamble transmissions can fail due to a collision or to a wireless channel error. That is, each one of the successful preambles has a certain probability of being correctly decoded at the eNB. The 3GPP suggests to model the probability of a wireless channel error in the kth preamble transmission of a UE as [1]

$$\Pr\left[\mathcal{E}_k\right] = 1/e^k \tag{3.26}$$

which decreases with k due to the power ramping process. This model has been adopted in most of the literature [30, 34, 109]. From there, we calculate the average preamble detection probability at the ith RAO as

$$f\left(i, \Pr\left[\mathcal{E}_k\right]\right) = \frac{1}{\mathbb{E}\left[N_i\right]} \sum_{k=1}^{k_{\max}} \left(1 - \Pr\left[\mathcal{E}_k\right]\right) \mathbb{E}\left[N_i\left(k\right)\right]. \tag{3.27}$$

Next, let $N_{D,i}$ be the RV that defines the number of UEs whose preamble transmissions are correctly decoded by the eNB at the ith RAO; its pmf is

$$p_{N_{D,i}}\left(s\right) = \sum_{v=s}^{r} \binom{v}{v-s} \left(1 - f\left(i, \Pr\left[\mathcal{E}_k\right]\right)\right)^{v-s} f\left(i, \Pr\left[\mathcal{E}_k\right]\right)^{s} p_{S_i}\left(v\right) \qquad \forall s \in \mathcal{S} \tag{3.28}$$

and its expected value is

$$\mathbb{E}\left[N_{D,i}\right] = \sum_{s=0}^{s_{\max}} s\, p_{N_{D,i}}\left(s\right). \tag{3.29}$$

At this point we have concluded the modeling of preamble transmission (*Msg1*), so we proceed to model the RAR (*Msg2*). Parameters involved in the RAR are the maximum number of available uplink grants per millisecond f_g and the length of the RAR window t_{rar}. With these values, we easily calculate the number of available uplink grants per RAR window as$g = f_g t_{\text{rar}}$. Common values for these parameters are listed in Table 3.1.

Let $\{N_{G,i}\}_{i\in\mathbb{N}}$ be the stochastic process that defines the number of UEs that will receive an uplink grant in response to a preamble transmitted at the ith RAO; its sample space is the number of successes $s \in \mathcal{S}$. In other words, $N_{G,i}$ is the RV that defines the number of UEs that successfully complete the first two steps of the RAP at the ith

RAR window. The pmf of RV $N_{G,i}$ is obtained by truncating the pmf of $N_{D,i}$ at $s = g$ and accumulating the remaining values in $\Pr[N_{G,i} = g]$. That is,

$$p_{N_{G,i}}(s) = \begin{cases} p_{N_{D,i}}(s), & \text{for } s \leq g-1 \\ \sum_{v=g}^{r} p_{N_{D,i}}(v), & \text{for } s = g. \end{cases} \tag{3.30}$$

Then we obtain $\mathbb{E}[N_{G,i}]$ analogously to (3.29). Next, let $N_{G,i}(k)$ be the RV that defines the number of UEs that successfully complete the first two steps of the RAP at the ith RAR window in the kth preamble transmission. The expected value of $N_{G,i}(k)$ is derived from that of $N_{G,i}$ as follows.

$$\mathbb{E}[N_{G,i}(k)] = \frac{\mathbb{E}[N_{G,i}]\ \mathbb{E}[N_i(k)]\ (1 - \Pr[\mathcal{E}_k])}{\mathbb{E}[N_i]\ f(i, \Pr[\mathcal{E}_k])}. \tag{3.31}$$

On the other hand, let $N_{F,i}(k)$ be the RV that defines the number of failed UE access attempts at the ith RAO; its expected value can be easily calculated from (3.31) as

$$\mathbb{E}[N_{F,i}(k)] = \mathbb{E}[N_i(k)] - \mathbb{E}[N_{G,i}(k)]. \tag{3.32}$$

It will be latter observed that the probability of failing the RAP during *Msg3* or *Msg4* transmissions is very close to zero. Therefore, $N_{F,i}(k)$ is sufficiently close to the exact number of UEs whose access attempt fails when considering the full RAP. Building on this, it is safe to assume $N_{F,i}(k)$ is the number of UEs whose kth access attempt fails at the ith RAO.

This concludes the analytical model of the RAR. But before moving into the model of the backoff procedure, we illustrate the difference between the pmfs of S_i, $N_{D,i}$, and $N_{G,i}$ for the $i = 343$th RAO under the TM 2 without the ACB scheme in Fig. 3.4. As mentioned in the previous chapter, this is a RAO of especial interest under the TM 2 without ACB scheme as is the first RAO in which the expected number of UE arrivals exceeds the capacity of the RAP (i.e., $\mathbb{E}[X_i] > \mathrm{C}(54, 15) = 15$). At this particular RAO, $\mathbb{E}[N_i] = 36.05$ and $\mathbb{E}[N_{G,i}] = 13.71$. Fig. 3.4 clearly shows that $\mathbb{E}[S_i] > \mathbb{E}[N_{D,i}] > \mathbb{E}[N_{G,i}]$, which showcases the great influence of g on the capacity of the RAP.

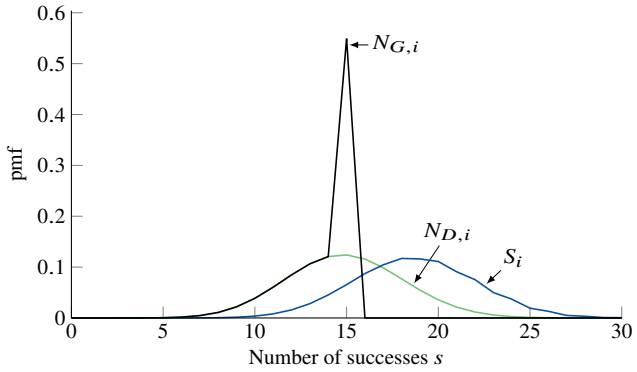


Figure 3.4: Pmf of the number of successful preambles S_i, decoded preambles at the eNB $N_{D,i}$, and assigned uplink grants $N_{G,i}$ for the $i = 343$th RAO under the TM 2.

Fig. 3.4 also allows us to showcase one of the main weaknesses of the analytical model proposed by Wei *et al.* [109]: the effect of g on the number of successful accesses. In their model, only the expected values of the previously introduced RVs are used. Therefore, the only possible approach to account for the effect of the number of available uplink grants g is to define $\mathbb{E}\left[N'_{G,i}\right] = \min\left\{\mathbb{E}\left[N_{D,i}\right], g\right\}$,[1] which is not accurate. For instance, at the 343th RAO this definition would lead to $\mathbb{E}\left[N'_{G,i}\right] = 15$ whereas the correct value is $\mathbb{E}\left[N_{G,i}\right] = 13.71$ as mentioned above. Now we proceed with the model of backoff time.

The UEs whose preamble transmission failed will not receive an uplink grant by the end of the next RAR window. Therefore, it is at the end of the RAR window that the UEs become aware of a failed access attempt. The time elapsed since preamble transmission and the end of the RAR window is given as

$$d_f = 1 + d_{\mathrm{p}} + t_{\mathrm{rar}}. \tag{3.33}$$

That is, one ms is required for preamble transmission, d_{p} ms to process the transmitted preambles at the eNB, and the duration of the RAR window is t_{rar} ms. If the maximum number of preamble transmissions $k_{\max}$ has not been reached, failed UEs:

[1]The apostrophe has been added to differentiate the definition of provided by Wei *et al.* [109] from ours.

1) increase the preamble transmission counter k; 2) wait for a random backoff time $t_b \equiv \mathrm{U}[0,1)\, b_{\max}$; and 3) transmit a newly selected preamble. The model for the backoff is described in the following.

Let B be the RV that represents the total number of RAOs a UE has to wait due to backoff during the RAP. Hereafter B is denoted simply as the backoff time, but is given in RAOs; hence it sample space is $i \in \mathbb{N}$. Also, let K be the RV that represents the number of preamble transmissions performed by a UE that successfully completes the RAP. Naturally, no backoff procedure is performed if a UE successfully completes the RAP in its first preamble transmission. Therefore, it is clear the pmf of $B \mid K = 1$ is

$$p_{B|K}(i \mid 1) = \delta(i)\,. \tag{3.34}$$

It is also clear that the conditional pmf of $B \mid K = 2$ is positive between $i_{B,\min} = \left\lceil d_f / t_{\mathrm{rao}} \right\rceil$ and $i_{B,\max} = \lceil (d_{\mathrm{f}} + b_{\max})/t_{\mathrm{rao}} \rceil$; hence,

$$p_{B|K}(i \mid 2) = \frac{1}{b_{\max}} \begin{cases} i\, t_{\mathrm{rao}} - d_{\mathrm{f}}, & \text{if } i = i_{B,\min} \\ t_{\mathrm{rao}}, & \text{if } i_{B,\min} < i < i_{B,\max} \\ d_f + b_{\max} - (i-1)\, t_{\mathrm{rao}}, & \text{if } i = i_{B,\max}. \end{cases} \tag{3.35}$$

The conditional pmf of $B \mid K = 2$ is of special importance because it allows us to model the backoff process at each RAO as follows.

Let $i_{\max}$ be the last RAO in which a preamble transmission can occur. Please also recall $y_{\max}$ is the maximum number of barring checks before determining an access failure and $i_{W,\max}$ is the maximum number of RAOs UEs can wait due to one failed barring check. With these data, $i_{\max}$ can be easily calculated as

$$i_{\max} = i_{\mathrm{dist}} + (k_{\max} - 1)\; i_{B,\max} + (y_{\max} - 1)\; i_{W,\max} \tag{3.36}$$

Next, we define $\imath_{\min} = \min\{i_{B,\min}, i\}$, $\imath_{\max} = \min\{i_{B,\max}, i\}$; these allow us to model the backoff process at each RAO by means of the following recursion

$$\mathbb{E}[N_i(k)] = \sum_{\nu=\imath_{\min}}^{\imath_{\max}} \mathbb{E}\left[N_{F,i-\nu}(k-1)\right]\, p_{B|K}(\nu \mid 2)\,,$$
$$\text{for } i = 1, 2, \ldots, i_{\max};\; k = 2, 3, \ldots, k_{\max} \tag{3.37}$$

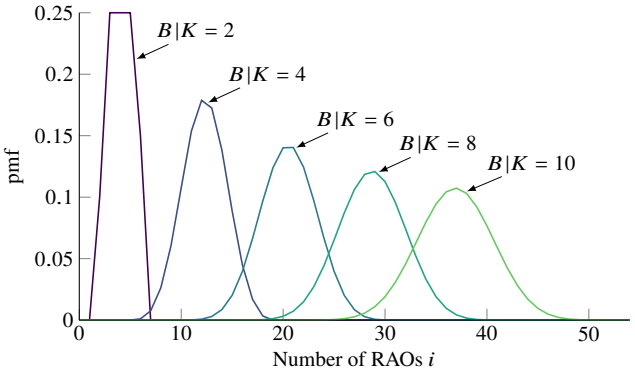


Figure 3.5: Pmf of the total number of RAOs a UE has to wait due to backoff given the UE succeeds at the kth preamble transmission $B \mid K \in \{2, 4, 6, 8, 10\}$ for $b_{\max} = 20$ ms.

given the initial condition $\mathbb{E}[N_1(k)] = 0$ for $k \geq 2$.

From (3.35), the pmf of $B \mid K$ can be calculated recursively as follows.

$$p_{B|K}(i \mid k) = \sum_{v=i_{B,\min}}^{i_{B,\max}} p_{B|K}(v \mid 2)\, p_{B|K}(i - v \mid k-1), \qquad \text{for } k = 3, 4, \ldots, k_{\max}. \tag{3.38}$$

Fig. 3.5 shows the pmf of the backoff time (in RAOs) given $t_{\text{rao}} = 5$ ms, $d_f = 8$ ms, and $b_{\max} = 20$ ms.

Finally, let D_{BO} the RV that represents the total delay due to backoff for a UE that successfully completes the RAP, given in milliseconds. Clearly, the pmf of $D_{\text{BO}} \mid K$ can be easily calculated as the function of $B \mid K$

$$\Pr[D_{\text{BO}} = i\, t_{\text{rao}} \mid K] = p_{B|K}(i \mid k). \tag{3.39}$$

The RAP concludes with the transmission of connection request (*Msg3*) and contention resolution (*Msg4*) messages. These are sent through dedicated resources, and are protected by robust a HARQ mechanism. Generally speaking, HARQ is a complex mechanism in which packets received with errors are stored instead of being discarded; these are combined with posterior retransmissions to increase the probability of recovering erroneous data. Hence, packets are not simply retransmitted in HARQ. Instead,

the amount of redundancy increases at each transmission, which also increases the probability of recovering the original packet.

Because of this complexity, the 3GPP simply advises to assume a constant error probability $\Pr[\mathcal{E}_h]$ to each of the *Msg3* and *Msg4* transmissions. A maximum number of transmissions for each of these packets $h_{\max}$ is also defined. Due to their similarity, we exemplify the modeling of both message transmissions with *Msg3* only as follows. The modeling for *Msg4* is analogous to that of *Msg3* but with a different RTT as shown in Table 2.3 on page 26 [3, Table 16.2.1-1].

Let D_{M3} be the RV that defines the time elapsed between the first *Msg3* transmission attempt by a given UE and its reception at the eNB, conditioned to the correct transmission of this message within the maximum number of attempts. The distribution of D_{M3} depends on the RTT d_{m3} ms, the error probability $\Pr[\mathcal{E}_h]$, and $h_{\max}$. Also let H be the RV that defines the number of attempts required for the successful transmission of *Msg3* whose support is $h \in \{1, 2, \ldots, h_{\max}\}$. It is easy to see that the pmf of $D_{\mathrm{M3}} \mid H = h$ is

$$p_{D_{\mathrm{M3}}|H}(d \mid h) = \delta(d - (h-1)\, d_{\mathrm{m3}}). \tag{3.40}$$

As described above, the probability of performing $h_{\max}$ attempts without success is extremely low even for relatively high values of $\Pr[\mathcal{E}_h]$. This holds true for the case in which both *Msg3* and *Msg4* are considered. Concretely, the probability that $h_{\max}$ *Msg3* or *Msg4* transmissions fail is

$$\Pr[\mathcal{E}_m] = \Pr[\mathcal{E}_h]^{h_{\max}} \left(2 - \Pr[\mathcal{E}_h]^{h_{\max}}\right) \tag{3.41}$$

which is very low (i.e., $2 \cdot 10^{-5}$) for the suggested value $\Pr[\mathcal{E}_h] = 0.1$ (see Table 3.1 on page 48). This value is coherent with the maximum admissible packet error ratio for data packets in LTE-A of 0.1 [4, Sec. 7.2.3]. Hence, $\Pr[\mathcal{E}_h] = 0.1$ can be seen as the worst case scenario for *Msg3* and *Msg4* transmissions, which takes us to the following proposition.

Proposition 3.3. *Given the probability of a UE failing an access attempt during the transmission of either* Msg3 *or* Msg4 *is extremely low, it is safe to assume these UEs do not go back to preamble transmission and terminate the RAP at this point.*

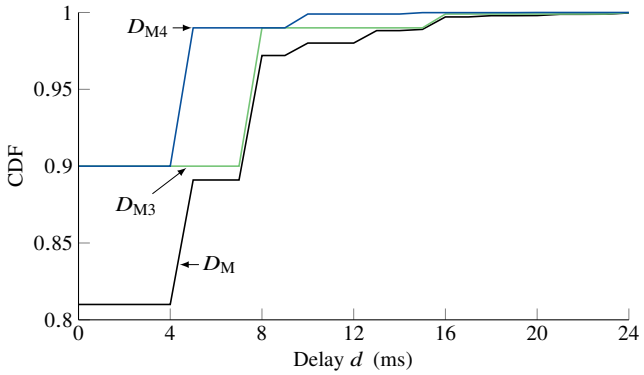


Figure 3.6: CDF of the access delay due to the transmission of *Msg3* and *Msg4* for the given error probability during transmission, $\Pr[\mathcal{E}_h] = 0.1$; the RTTs of *Msg3* and *Msg4* are 8 and 5 ms, respectively.

Building on this proposition, the distribution of D_{M3} alone can be calculated as follows.

$$p_{D_{\mathrm{M3}}}(d) = \frac{1 - p_{\mathcal{E}_h}}{1 - p_{\mathcal{E}_h}^{h_{\max}}} \sum_{h=1}^{h_{\max}} p_{\mathcal{E}_h}^{h-1} \, \delta\left(d - (h-1)\, d_{\mathrm{m3}}\right) \tag{3.42}$$

where $1 - p_{\mathcal{E}_h}^{h_{\max}}$ is a normalization factor to ensure $p_{D_{\mathrm{M3}}}(d)$ is a pmf.

As stated above, the pmf of D_{M4} can be obtained in analogously to that of D_{M3} but substituting the RTT, d_{m3}, with d_{m4}.

Next, let D_{M} be the RV that denotes the time elapsed between the first transmission attempt of *Msg3* and the successful transmission of *Msg4*. Naturally, D_{M} is the sum of RVs D_{M3} and D_{M4}, whose pmf can be calculated by the convolution

$$p_{D_{\mathrm{M}}}(d) = \Pr[D_{\mathrm{M3}} + D_{\mathrm{M4}} = d] = \sum_{v=0}^{d} p_{D_{\mathrm{M3}}}(v) \; p_{D_{\mathrm{M4}}}(d - v) \tag{3.43}$$

since $D_{\mathrm{M3}} \perp D_{\mathrm{M4}}$. Fig. 3.6 shows the CDF of D_{M3}, D_{M4}, and D_{M} for the selected configuration.

Now we are able to calculate the number of UEs that successfully complete the RAP. For this, let the RV $N_{S,i}(k)$ define the number of UEs that successfully complete

the RAP in the kth preamble transmission, performed at the ith RAO. The expected value of $N_{S,i}(k)$ can be calculated as

$$\mathbb{E}\left[N_{S,i}(k)\right] = (1 - \Pr[\mathcal{E}_m])\, \mathbb{E}\left[N_{G,i}(k)\right]. \tag{3.44}$$

The final component of our model is the calculation of the minimum time needed to complete the RAP

$$d_{\min} = \min\{d \mid \Pr[D = d] \geq 0\} = 4 + d_{\mathrm{p}} + d_{\mathrm{ug}} + d_{\mathrm{cr}}. \tag{3.45}$$

That is, 4 ms are needed for the transmission of the four messages that comprise the RAP; d_{p}, d_{ug}, and d_{cr} are the processing delays of the preamble, uplink grant, and connection request messages, respectively [3, Table 16.2.1-1]. From there, we can define the RV $D_{\min}$ as the minimum access delay, whose pmf is

$$p_{D_{\min}}(d) = \delta(d - d_{\min}). \tag{3.46}$$

3.3.4 Obtaining the KPIs

In this brief subsection we describe the process for obtaining the KPIs defined in Section 3.1 for the performance evaluation of the RAP.

The first and most important KPI is the success probability, denoted as P_s and defined as the ratio of successful to total UEs. Please recall that RV $N_{S,i}(k)$ defines the number of successful accesses that occur at the kth preamble transmission and at the ith RAO. Building on this, let N_S be the RV that defines the total number of successful accesses throughout the access period, its expected value is

$$\mathbb{E}[N_S] = \sum_{i=0}^{i_{\max}} \sum_{k=1}^{k_{\max}} \mathbb{E}\left[N_{S,i}(k)\right]. \tag{3.47}$$

Hence, P_s is simply

$$P_s = \frac{\mathbb{E}[N_S]}{n}. \tag{3.48}$$

Next, the collision probability, denoted as P_c and defined as the ratio of collided to available preambles throughout the access period is calculated as

$$P_c = \frac{1}{(i_{\max}+1)\, r} \sum_{i=0}^{i_{\max}} \sum_{c=1}^{r} c p_{C_i}(c) \tag{3.49}$$

where C_i defines the number of collided preambles at the ith RAO.

The pmf of RV K, which defines the number of preamble transmissions performed by the successfully accesses UEs, is calculated as follows.

$$p_K(k) = \frac{1}{\mathbb{E}[N_S]} \sum_{i=0}^{i_{\max}} \mathbb{E}\left[N_{S,i}(k)\right], \qquad \text{for } k = 1, 2, \ldots, k_{\max}, \tag{3.50}$$

From there, its expected value $\mathbb{E}[K]$ and ϕth percentile K_ϕ can be calculated. The former is given analogously to (3.29) and the latter is derived by means of a linear interpolation of the CDF of K $F_K(k)$.

The calculation of the access delay D concludes our model. But first, the delay induced by the transmission of RAR messages within the RAR window, defined by RV D_{RAR}, must be calculated. for this, please recall that there are f_g uplink grants available per millisecond (i.e., per subframe), and that the length of the RAR window is t_{rar} ms. Building on this, the sample space for D_{RAR} is $d \in \{0, 1, \ldots, t_{\text{rar}}\}$ ms and its pmf can be obtained as

$$p_{D_{\text{RAR}}}(d) = \frac{1}{\mathbb{E}[N_S]} \sum_{i=0}^{i_{\max}} \max\left\{0, \min\left\{f_g t_s, \mathbb{E}[M_S(i)] - (d\, f_g)\right\}\right\}; \qquad \text{for } d = 0, 1, \ldots t_{\text{rar}} - 1. \tag{3.51}$$

Finally, the pmf of D can be calculated as

$$p_D(d) = \Pr\left[D_{\text{ACB}} + D_{\text{BO}} + D_{\text{RAR}} + D_{\text{M}} + D_{\min} = d\right]. \tag{3.52}$$

That is, we assume independence between the RVs that contribute to the access delay. These are, from left to right in (3.52) the increase in delay due to: 1) the ACB scheme D_{ACB}; 2) backoff D_{BO}; 3) RAR D_{RAR}; and 4) *Msg3* and *Msg4* transmissions D_{M}.

The last component of (3.52) is the minimum time needed to complete the RAP. The expected value of the access delay $\mathbb{E}[D]$, its CDF $F_D(d)$, and ϕth percentile D_ϕ are obtained analogously to those of K.

3.3.5 Assessing the accuracy of our model

In the following section, we assess the accuracy of our model with respect to simulations. Similarly as in Chapter 2, each simulation starts at $i = 0$ and ends when every UE has terminated the RAP and the number of simulation runs is set to the smallest number that ensures that all the cumulative KPIs obtained up to the last simulation differ from those obtained up to the previous simulation by less than 0.01 percent. Then, we compare the accuracy of our model with that of the RM [109]. For this, we first define adequate metrics to assess the accuracy of the obtained KPIs.

It is clear that the accuracy of the success P_s and collision P_c probabilities can be easily assessed in terms of the relative error. On the other hand, assessing the accuracy of the pmfs of the number of preamble transmissions and access delay requires more sophisticated mechanisms. In our previous work [60, 69], the accuracy of these pmfs was assessed in terms of the relative error of numerous percentiles. Nevertheless, this mechanism is not capable of reflecting the level of likeliness between the pmfs obtained by simulation and by an analytical model.

Building on this, we now assess the accuracy of our model and of the RM by means of the Jensen-Shannon Divergence (JSD). The JSD measures the increase in the Shannon's entropy when a given pmf is assumed to be the real pmf of an RV. In other words, the JSD measures the loss of information when an approximated pmf is assumed to be the real pmf of an RV. It is defined for two pmfs as follows.

$$\mathrm{JSD}\left(p_{\mathrm{sim}}(x), p(x)\right) \equiv H\left(\frac{p_{\mathrm{sim}}(x) + p(x)}{2}\right) - \frac{H\left(p_{\mathrm{sim}}(x)\right) + H\left(p(x)\right)}{2}. \tag{3.53}$$

where $p_{\mathrm{sim}}(x)$ is the real pmf (in our case, the one obtained by simulation), $p(x)$ is the pmf we assume to be the real one, and $H(\cdot)$ is the base-e Shannon's entropy defined as

$$H\left(p(x)\right) \equiv -\sum_{\forall x \in X} p(x) \log p(x). \tag{3.54}$$

When the base-e Shannon's entropy is used, we have

$$0 \leq \mathrm{JSD}\left(p_{\mathrm{sim}}(x), p(x)\right) \leq \log 2, \tag{3.55}$$

where JSD = 0 indicates that both pmfs are identical. On the other hand, JSD = $\log 2$ can only be obtained if no element in the support of one of the RVs is in the support of the other.

3.4 Results and discussion

In this section we showcase the accuracy of our model by presenting the results obtained from the performance evaluation of the RA in cellular networks. Concretely, we compare the accuracy of our model with that of the model presented by Wei *et al.* [109] with respect to simulations. As mentioned above, we refer to the model presented by Wei *et al.* simply as the RM, and was the most thorough and accurate analytical model of the RAP prior to ours. The PRACH and PDCCH configuration parameters are enlisted in Table 3.1 on page 48.

The performance evaluation of the RA was conducted under two mMTC scenarios in which $n = 30\,000$ UE arrivals occur according to TM 2. As such, this section is divided in two subsections. In the first one, we present our results given no ACB scheme is implemented. In the second one, we present our results given the ACB scheme is implemented with $p_{\mathrm{acb}} = 0.5$ and $t_{\mathrm{acb}} = 4$; these are fixed throughout the access period. As observed in Chapter 2, this is the combination of barring parameters, among those available for selection at the SIB 2, that leads $P_s \geq 0.95$ with the minimum D_{95} [68, 95].

As described above, our model was designed under the assumption that preambles transmitted by multiple UEs are never decoded by the eNB. Therefore, the vast majority of the results presented in this section were obtained under this assumption. In addition, our model was modified to accommodate the assumption that preambles transmitted by multiple UEs are always decoded by the eNB. Results obtained under this assumption are presented at the end of this section.

3.4.1 Disabled ACB scheme

As a starting point, we show the expected number of successful accesses $\mathbb{E}[N_{S,i}]$ at each RAO obtained by simulation, by the RM, and by our model Fig. 3.7a. In addition, Fig. 3.7b shows the absolute error obtained by calculating $\mathbb{E}[N_{S,i}]$ with either of both analytical models with respect to simulations.

Fig. 3.7 clearly shows that the results obtained by both models and by simulation are extremely similar for most of the RAOs. The most notorious exception is observed for the RM in RAOs where $\mathbb{E}[N_{S,i}] \approx 15$; here, an absolute error of up to 2 successful accesses is observed. The main reason for this is that the expected number of assigned uplink grants per RAO is calculated directly from the expected number of decoded preambles. As a result, the RM overestimates the number of successful accesses for the selected scenario. This problem has also been identified by Arouk *et al.* [20].

In our model, we follow a different approach and use the whole pmf (see (3.30) and Fig. 3.4). As a result, this error is not present in our model, which leads to a maximum absolute error of up to one order of magnitude lower than the obtained with the RM as illustrated in Fig. 3.7.

To proceed, we show the KPIs obtained by simulation for $t_{\text{rao}} \in \{2, 5, 10\}$ ms in Table 3.2; these values correspond to PRACH configurations *prach-ConfigIndex* $\in \{3, 6, 13\}$, respectively.

Now we provide with an in-depth look at the accuracy of the RM and of our proposed model by listing the relative error of these models with respect to simulations for the KPIs listed in Table 3.2. It is clear that the higher accuracy of our model is reflected in the obtained KPIs, as the relative error is less than 3 percent for all the values shown in Table 3.2. For instance, the relative error exceeds 2 percent only twice with our model, which accounts for less than 6 percent of the listed KPIs. In contrast, the RM leads a relative error higher than 2 percent for more than half of the listed KPIs and the error is particularly high for the access delay. One of the main contributing factors to such a high error is the use of the expected delay of RAR, *Msg3*, and *Msg4* transmissions instead of the pmf. This causes the abrupt changes in the CDF of access delay calculated with the RM, depicted in Fig. 3.8. It can also be seen in Fig. 3.8 that

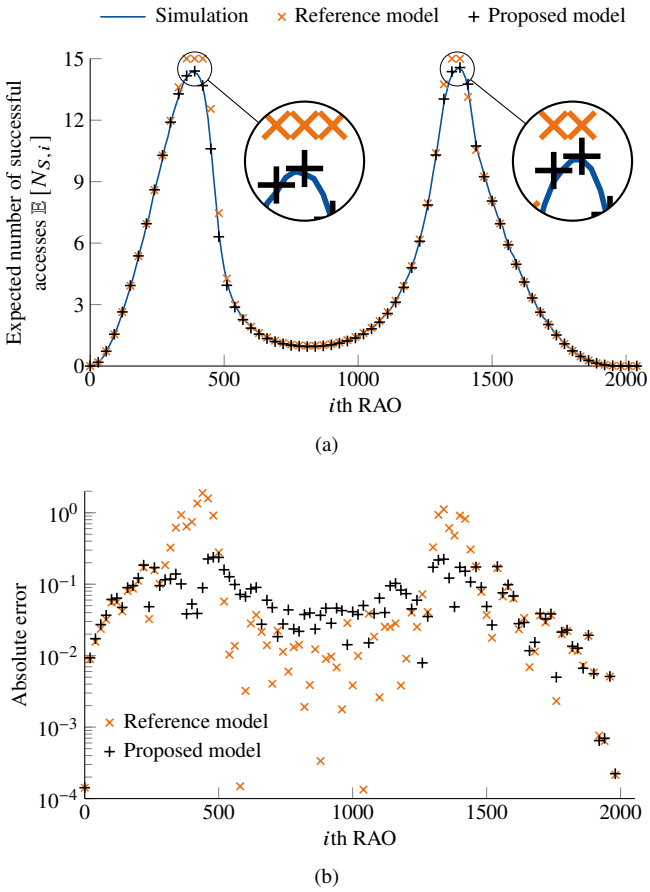


Figure 3.7: (a) Comparison and (b) absolute error (in logarithmic scale) of the expected number of successful accesses $\mathbb{E}[N_{S,i}]$ at each RAO obtained by simulation, by the RM [109], and by our proposed model; disabled ACB scheme.

the CDF of access delay calculated with our model highly resembles the one obtained by simulation.

To further showcase the advantages of our model with respect to the RM, we show the JSD of both models with respect to simulations in Fig. 3.9 for RVs K and D.

Table 3.2: KPIs obtained by simulation with different values of t_{rao} (ms); disabled ACB scheme.

KPI	$t_{\text{rao}} = 2$	$t_{\text{rao}} = 5$	$t_{\text{rao}} = 10$
Success probability (%)	66.44	31.33	9.89
Collision probability (%)	18.02	43.48	60.02
Number of preamble transmissions			
Expected value	4.10	3.45	3.23
10th percentile	1.00	1.00	1.00
50th percentile	2.87	1.98	1.83
90th percentile	8.01	7.30	6.93
95th percentile	8.95	8.57	8.38
Access delay (ms)			
Expected value	67.52	68.76	81.42
10th percentile	15.00	15.08	15.01
50th percentile	54.05	46.93	54.16
90th percentile	136.66	155.60	196.35
95th percentile	153.52	182.59	236.56

As described in Section 3.3, the JSD is a measure of the loss of information about a RV when an approximated pmf is assumed to be the real one. The pmf obtained by simulation is assumed to be the real pmf and the base-e logarithm was used, so the JSD is lower bounded by 0 and upper bounded by $\log 2$.

Fig. 3.9 clearly shows that the pmf of K obtained with both models is highly similar to the one obtained by simulation. For instance, the JSD of our model decreases slightly as t_{rao} increases. On the other hand, the JSD of the RM for RV K presents a drastic drop for $t_{\text{rao}} = 10$. The reason for this phenomenon can be inferred from Fig. 3.7b: the accuracy of the RM increases as the number of successful accesses decreases during periods of high congestion and $t_{\text{rao}} = 10$ is most congested scenario of the three. At first glance this seems to be an advantage of the RM over ours, but this increase in the

Table 3.3: Relative error (%) for the reference model (RM) and our proposed model (PM) with different values of t_{rao} (ms); no ACB scheme.

	$t_{\text{rao}} = 2$		$t_{\text{rao}} = 5$		$t_{\text{rao}} = 10$	
KPI	RM	PM	RM	PM	RM	PM
Success probability	0.35	0.08	2.70	0.29	0.59	0.36
Collision probability	1.15	0.13	1.63	0.20	0.02	0.09
Number of preamble transmissions						
Expected value	1.39	0.69	2.90	0.97	0.35	0.76
10th percentile	0.00	0.00	0.00	0.00	0.00	0.00
50th percentile	3.19	1.54	3.60	1.71	0.48	1.29
90th percentile	0.28	0.08	2.35	0.48	0.22	0.50
95th percentile	0.11	0.02	1.07	0.20	0.20	0.16
Access delay						
Expected value	21.36	2.17	3.18	2.59	11.53	2.28
10th percentile	26.67	0.10	25.97	0.55	59.92	0.30
50th percentile	10.59	1.64	12.82	2.30	0.59	2.16
90th percentile	19.26	0.12	10.62	0.33	2.34	0.38
95th percentile	19.17	0.09	6.51	0.34	10.08	0.24

accuracy of K and in P_c (see Table 3.3) are not reflected on P_s nor in D. In fact, the 10th percentile of D obtained with the RM given $t_{\text{rao}} = 10$ ms results in the highest relative error overall, exceeding 50 percent.

In fact, the JSD of the RM for RV D is so high that is comparable to $\log 2$; this is the upper bound for the JSD. For instance, the JSD obtained by comparing a uniform distribution $\mathrm{U}[0, 400]$ to the pmf of D obtained by simulation given $t_{\text{rao}} = 5$ is even lower: JSD = 0.304. In other words, more information about D is lost by using the RM than by assuming D is uniformly distributed between 0 and 400. The JSD for D obtained with our model is more than two orders of magnitude lower.

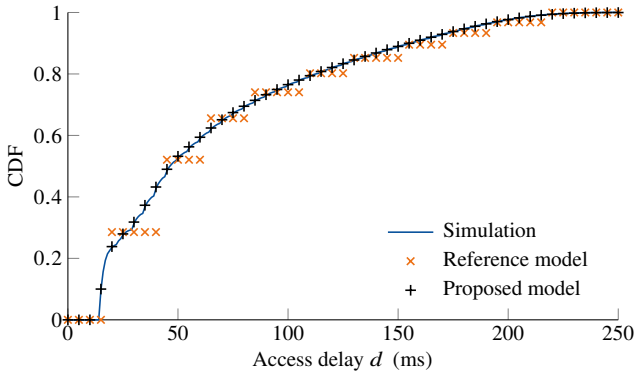


Figure 3.8: CDF of access delay $F_D(d)$; disabled ACB scheme.

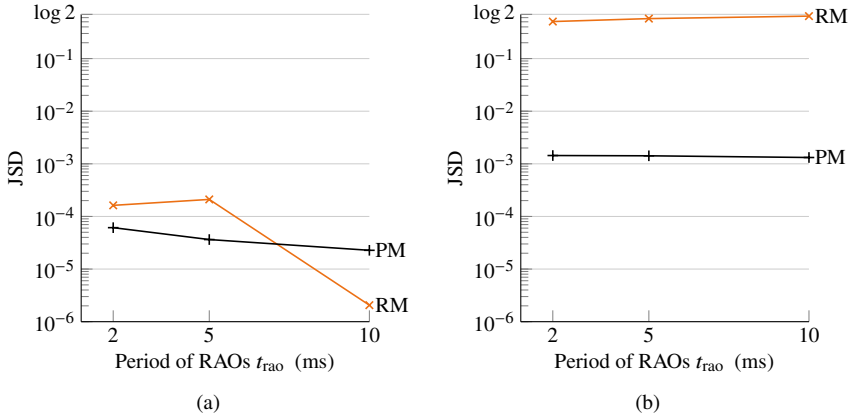


Figure 3.9: JSD in the pmfs of (a) the number of preamble transmissions K and (b) access delay D obtained by simulation and by the analytical models; disabled ACB scheme.

3.4.2 Enabled ACB scheme

Now we preset our results when the ACB scheme is implemented. We have observed previously that the accuracy of our model not affected by different values of t_{rao}.

Therefore, from this point on, results are only presented for $t_{\text{rao}} = 5$.

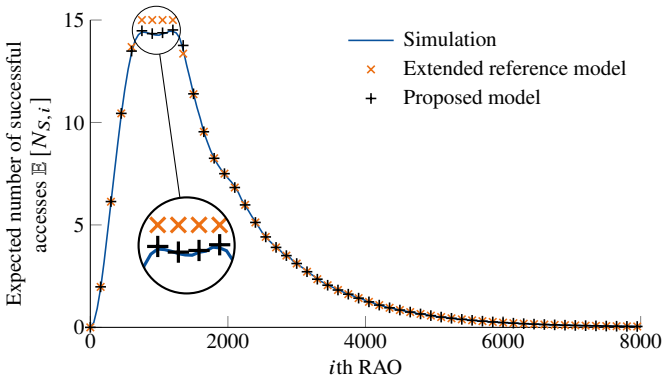
It is important to mention that the ACB scheme was not included in the RM. Therefore, we incorporated our model of the ACB scheme as an input to the RM. Then, we calculate the overall access delay D as the convolution of the pmf of access delay provided by the RM and the pmf of the barring time D_{ACB}. By doing so, we are able to obtain the desired results with a model that only comprises the RAP. We denote this extension as the extended RM (ERM).

Once more, our evaluation begins with the expected number of successful accesses at each RAO $\mathbb{E}[N_{S,i}]$. These are shown in Fig. 3.10a, and the absolute error for $\mathbb{E}[N_{S,i}]$ between the analytic models and simulations is shown in Fig. 3.10b. Here we observe the same behavior described above: the RM overestimates $\mathbb{E}[N_{S,i}]$ for values close to g. On the other hand, the absolute error obtained with our model is always below 0.1, except at one RAO, where it is lower than 0.2.

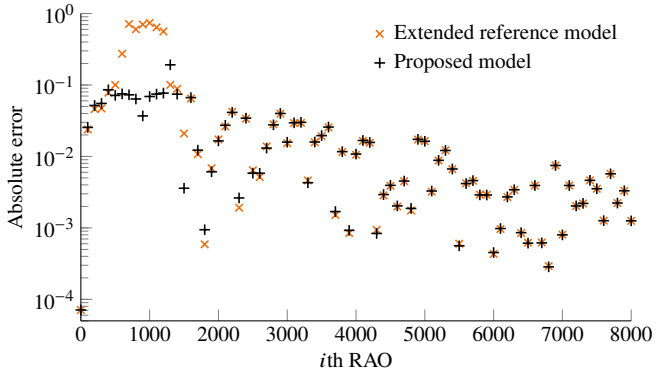
Results shown in Fig. 3.10 illustrate the accuracy of our model of the RAP, but also that of the ACB scheme. That is, the highest error marks observed in Fig. 3.10b are due to the RM itself.

Table 3.4 shows the KPIs obtained by simulation and the relative error of both analytic models. Clearly, our model is much more accurate than the ERM. The only two exceptions are in the expected and the 10th percentile of access delay. The reason for this is that we assume the RVs of delay at each of the steps of RA are independent, which is not the case. For instance, the UEs whose access is barred a relatively long time are more likely to contend with fewer UEs. On the other hand, UEs that succeed in the first barring check are more likely to contend with a large number of UEs; hence, preambles transmitted by these UEs have a higher collision probability than those transmitted by highly delayed UEs. Nevertheless, this effect fades when high percentiles of delay are considered and our model provides more information on the real pmf of D than the ERM. To support this statement, we present the JSD of both models with respect to simulations in Fig. 3.11 for RVs K and D.

Fig. 3.11 clearly shows that the JSD of our model is more than one order of magnitude lower than that of the ERM for both RVs even though some values presented in Table 3.4 suggest a slightly higher accuracy for the ERM. In other words, it is mere



(a)



(b)

Figure 3.10: (a) Comparison and (b) absolute error of the expected number of successful accesses at each RAO $\mathbb{E}[N_{S,i}]$, obtained by simulation, by the ERM, and by our proposed model; implemented ACB scheme with fixed $p_{\text{acb}} = 0.5$ and $t_{\text{acb}} = 4$.

coincidence that some relative errors obtained with the ERM are lower than those obtained with our model. In other words, the JSD provides more information on the accuracy of the pmfs obtained by the analytical models.

We provide a close look at the behavior of the CDF of D in Fig. 3.12. Specifically, Fig. 3.12a shows the complete CDF of D, then Fig. 3.12b shows only its first 250 ms.

Table 3.4: KPIs obtained by simulation and the relative error obtained by the ERM and by our proposed model (PM) for the selected scenario; ACB scheme with fixed $p_{\text{acb}} = 0.5$ and $t_{\text{acb}} = 4$.

		Rel. error (%)	
KPI	Simulation	ERM	PM
Success probability	97.48 %	1.39	0.18
Collision probability	1.62 %	18.00	3.28
Number of preamble transmissions			
Expected value	2.45	7.29	1.35
10th percentile	1.00	0.00	0.00
50th percentile	1.40	5.77	2.04
90th percentile	4.54	14.31	1.70
95th percentile	6.13	13.30	1.40
Access delay			
Expected value	4141.86 ms	2.36	2.37
10th percentile	18.12 ms	4.83	12.32
50th percentile	2945.89 ms	92.63	4.47
90th percentile	11 839.26 ms	1.04	1.04
95th percentile	15 809.89 ms	0.88	0.87

Clearly, the CDF obtained by the analytical models grows more rapidly than the one obtained by simulation. This phenomenon, combined with the fact that the ERM overestimates the number of successful accesses, causes a relative error of 92.63 percent in the 50th percentile of access delay obtained with the latter. However, the main reason for such a high error in this particular percentile is that it coincides with the barring rate $p_{\text{acb}} = 0.5$, which causes the step observed at the first RAOs of Fig. 3.12a. On the other hand, it is observed in Fig. 3.12b that this error is relatively low, but causes the CDFs obtained by the ERM to exceed 0.5 much earlier than the one obtained by simulation.

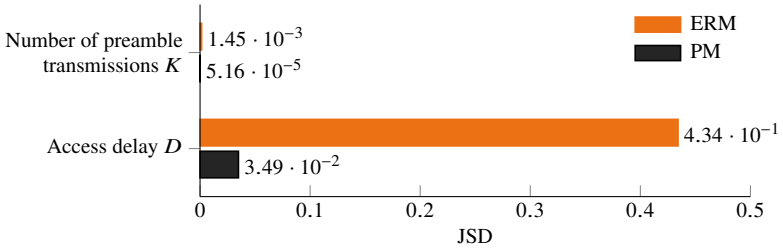
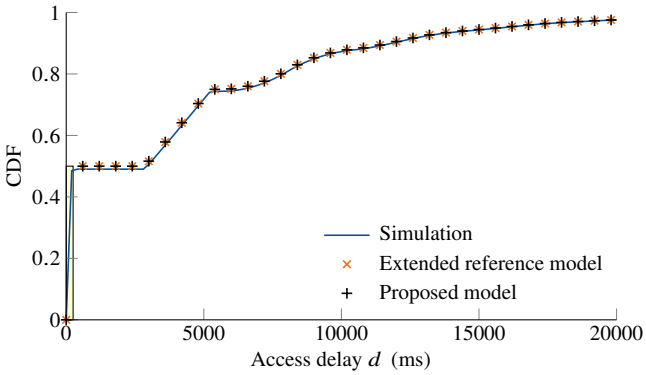


Figure 3.11: JSD in the pmfs of the number of preamble transmissions K and access delay D obtained by simulation and by the analytical models; implemented ACB scheme with fixed $p_{\text{acb}} = 0.5$ and $t_{\text{acb}} = 4$.

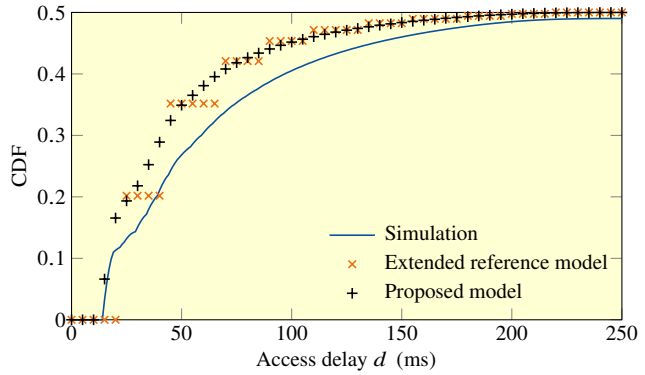
3.4.3 The eNB decodes the preambles transmitted by multiple UEs

As briefly explained in Section 3.2, two outcomes are possible when multiple UEs transmit the same preamble at the same RAO. In the first one, the eNB does not decode these preambles; throughout this study, we have assumed that this outcome occurs every time multiple UEs transmit the same preamble. In the second one, the eNB correctly decodes the transmitted preambles. The two major effects of decoding the preambles transmitted by multiple UEs are: 1) uplink grants may be sent in response to preambles transmitted by multiple UEs and 2) the multiple UEs that receive the same uplink grant will send their *Msg3*s in the same reserved uplink resources. Needless to say, these two effects negatively impact the performance of the RAP [34, 79, 95], and their implications are described in the following.

If uplink grants are sent in response to preambles transmitted by multiple UEs, less than g uplink grants will be available to respond to successful preambles. In other words, downlink resources are wasted on UEs that have cannot successfully completing the RAP. Furthermore, if multiple UEs send their *Msg3* in the same reserved uplink resources, a collision will occur at this point. Moreover, the UEs will not be aware of the collision until h_{max} *Msg3*s are transmitted and no *Msg4* is received. That is, only after the maximum number of *Msg3* transmissions is reached; only then, these UEs will perform backoff. As a consequence, the delay of these UEs will increase when compared to that of the outcome assumed throughout this thesis.



(a)



(b)

Figure 3.12: (a) Overall view and (b) first 250 ms [colored area in the lower left corner of (a)] of the CDF of the access delay $F_D(d)$ obtained by simulation, by the ERM and by our model; implemented ACB scheme with fixed $p_{\text{acb}} = 0.5$ and $t_{\text{acb}} = 4$.

We have adapted our model and also the RM, also referred to as the ERM, in order to evaluate the performance of the RAP when uplink grants may be transmitted in response to collided preambles. Table 3.5 shows the results obtained by simulation, by adapting the RM and by our model. The same principles employed to adapt our model were used to adapt the RM.

Fig. 3.13 shows the $\mathbb{E}[N_{S,i}]$ obtained by simulation and by both analytical models, along with the absolute error when compared to simulations. It can be easily observed from Fig. 3.13a that the achieved $\mathbb{E}[N_{S,i}]$ is lower under this assumption than in the previous one, but also that both analytical models exhibit a roughly similar accuracy. This is confirmed in Fig. 3.13b, where the absolute error is always below 1. The reasons for this similar accuracy are twofold. First, $\mathbb{E}[N_{S,i}]$ is nowhere near g; hence the accuracy of the ERM is not greatly affected. Second, the adaptations included in the analytical models are analogous and have a great impact on performance.

Next, we present the KPIs obtained by simulation and the relative errors obtained with the analytical models. As it can be seen, our model exceeds the accuracy of the ERM in success probability, collision probability and several delay percentiles. The accuracy of both models is similar for the different metrics of the number of preamble transmissions. We have also observed that the accuracy of both is similar by comparing the JSD and this similarity holds when the ACB scheme is implemented.

3.5 Conclusions

This chapter presented a thorough analytic model of the RA in cellular networks that includes the model of the ACB scheme with fixed parameters. This model was used to accurately assess the performance of RA under mMTC scenarios. Specifically, the accuracy of our model was assessed under several PRACH configurations with respect to simulation results and then compared it with that of the model proposed by Wei *et al.* Although the latter was the most accurate model prior to ours, its accuracy drops when the number of successful accesses per RAO approximates the RAP capacity. That is, when most of the available resources are being utilized. These are the scenarios of highest interest because the main objective of access control schemes is that of reducing congestion while maximizing the utilization of resources. This latter statement will become evident in the next chapter, in which our adaptive solution to support mMTC is presented.

Throughout this chapter, it was observed that the accuracy of our model is not affected by the distribution of the UE arrivals, by the signaling traffic intensity, nor

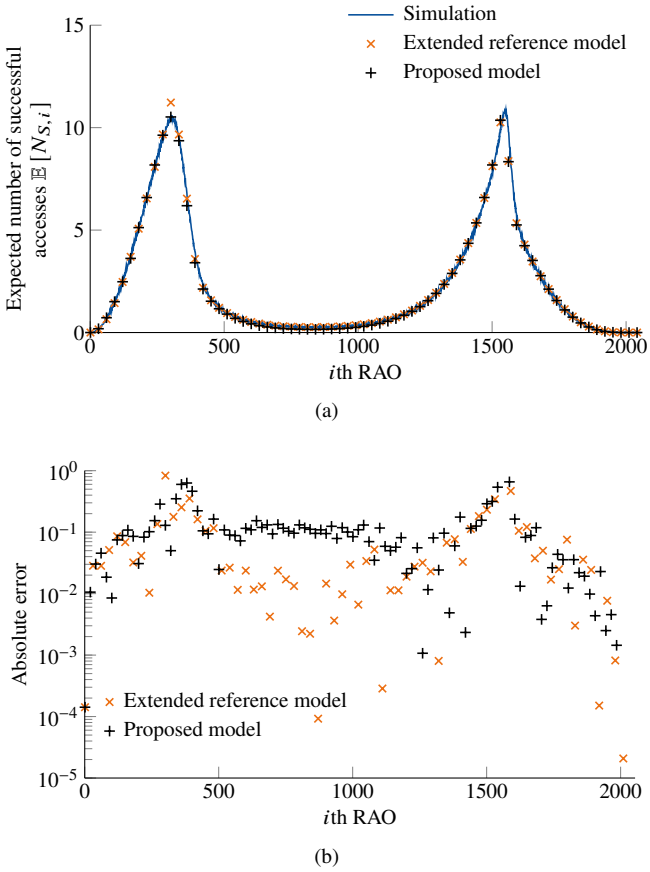


Figure 3.13: (a) Comparison and (b) absolute error of the expected number of successful accesses at each RAO $\mathbb{E}[N_{S,i}]$, obtained by simulation, by the ERM, and by our proposed model; the eNB decodes collided preambles.

by the selected PRACH configuration; still, it maintains an acceptable degree of computational complexity. For instance, by implementing our model in Octave, results were obtained within a few tens of seconds for the case in which no ACB is implemented and within a couple of minutes for the case in which the static ACB is implemented.

In addition, we adapted our model in order to evaluate the performance of the

Table 3.5: KPIs obtained by simulation and the relative error obtained by the ERM and by our proposed model (PM) for the selected scenario; the eNB decodes the preambles transmitted by multiple UEs .

KPI	Simulation	Rel. error (%)	
		ERM	PM
Success probability	16.42 %	1.36	0.30
Collision probability	49.46 %	0.38	0.09
Number of preamble transmissions			
Expected value	3.43	1.42	1.48
10th percentile	1.00	0.00	0.00
50th percentile	1.88	2.35	2.66
90th percentile	7.53	0.66	0.65
95th percentile	8.73	0.28	0.28
Access delay			
Expected value	103.38 ms	3.57	8.27
10th percentile	15.00 ms	26.67	3.95
50th percentile	69.83 ms	24.34	19.23
90th percentile	256.60 ms	0.44	5.97
95th percentile	306.21 ms	5.39	5.33

RAP under the assumption that the eNB correctly decodes the preambles transmitted by multiple UEs. The same process used to adapt our model was used to adapt the reference model in order to consider this latter assumption. Results show that the accuracy of our model is preserved even under this assumption.

A distinctive characteristic of our model is the inclusion of the static ACB scheme. Results indicate that our model of the ACB scheme is highly accurate and can be easily incorporated into other analytic models. For instance, it was incorporated to the model provided by Wei *et al.* to obtain the results presented in Table 3.4. However, there are a two main considerations that must be taken into account when incorporating our model of the ACB scheme to other models for the RAP. The first one is that percentiles

of access delay that are close to the selected barring rate may be highly affected by the lack of accuracy of the selected model of the RA procedure. The second one is that the CDF of access delay obtained by our model may raise more rapidly in the first few subframes than the one obtained by simulation. The reason for this is that we assume the RVs involved in the calculation of the access delay are independent, while a slight correlation exists. However, doing otherwise would greatly complicate our model, which in turn would compromise its computational efficiency.

Chapter 4

Adaptive access control for efficient mMTC in cellular networks

4.1 Introduction

Chapters 2 and 3 have focused on studying the efficacy of the RA procedure (RAP) defined by the 3rd Generation Partnership Project (3GPP) for cellular networks under massive machine-type communication (mMTC) applications. Results shown in both of these chapters have revealed that an access control scheme is needed under these conditions to prevent congestion in the random access (RA) channels. One of the most promising access control schemes is the access class barring (ACB), which redistributes the access of user equipments (UEs) through time to reduce the signaling traffic intensity. In particular, under the ACB scheme, each UE begin the RAP with probability equal to the barring rate. Otherwise, the UE must wait for a random time, selected according to the mean barring time. Both barring parameters, the barring rate and mean barring time are provided by the cellular base station (known as the evolved NodeB (eNB) in 4th generation (4G)).

A traditional ACB scheme in which the barring parameters are fixed throughout the operation of the network has also been studied in Chapter 2 and Chapter 3. In this chapter, we refer to the methods to select the barring parameters as the access class

barring configuration (ACBC) schemes. Hence, we refer to the traditional approach to use fixed parameters as the fixed ACBC scheme. Results presented in previous chapters highlight that the fixed ACBC scheme can relieve congestion under highly synchronized mMTC applications given adequate barring parameters are selected. Nevertheless, the fixed ACBC scheme presents two main drawbacks:

1. The barring parameters must be selected before the distribution period of UEs. Hence, no prior knowledge of the number of accessing UEs nor of the distribution of accesses may be available. If barring parameters are not selected to suit these two characteristics, severe congestion will occur.

2. The access of UEs will be delayed even under low signaling traffic scenarios (i.e., when no congestion occurs). Clearly, this is not desirable.

Efficiently supporting mMTC is one of the pillars of 5th generation (5G) [6, 49], and 4G will surely be the basis of 5G networks. Therefore, recent research efforts have been focused on ACBC schemes to adapt the barring parameters to the intensity of accesses in real time [34, 35, 71, 94, 97]. In theory, such ACBC schemes are the only ones capable of providing an optimal performance. That is, maximize the utilization of resources and to guarantee the access of the vast majority of the UEs to the eNB.

However, the vast majority of ACBC schemes presented in the literature [35, 71, 94] were designed for an idealized ACB scheme, in which every UE is subject to the ACB scheme even after the beginning of its RAP and in which there is no delay in the notification mechanisms (i.e., System Information Block (SIB) 2 transmissions). Therefore, these cannot be implemented at the eNB in their current form.

The development of a dynamic ACBC scheme capable of adapting its parameters in a real-time fashion is a challenging task [34] that is mainly hindered by the following factors.

1. The selectivity of the ACB scheme: Only the UEs that have not yet begun the RAP are subject to the ACB scheme. That is, those who have not yet performed their first preamble transmission. Once a UE has transmitted its first preamble, it is no longer subject to the ACB scheme [10, 53].

2. The delay of notification mechanisms: The barring parameters are broadcast through the SIB 2, whose period is much longer than the period of random access opportunities (RAOs) (resources in which preamble transmissions are allowed). Therefore, it is not possible to set precise barring parameters in a RAO-by-RAO basis.

3. The limited information available at the eNB regarding the number of contending UEs: The eNB ignores the number of contending UEs at a given time, the exact number of UEs deployed within the cell and, clearly, the distribution that the UE accesses will follow before these occur. After each RAO, the eNB is clearly aware of the number of successful accesses, but the number of failed accesses may not be known as several causes for an access failure exist; these were explained in detail on page 44 (Chapter 3). Therefore, the eNB can only approximate the number of UEs deployed in a cell based on the number of UEs registered previously and the number of decoded preambles at each RAO. Needless to say, the accuracy of such an approximation will suffer.

In this chapter, we present an adaptive ACBC scheme that can be directly implemented in the current 4G and in the coming 5G systems. Our ACBC scheme relies on the number of UEs that successfully complete the RAP and in an adaptive filtering process to adjust the barring parameters to the perceived signaling traffic intensity. The main objective of the filtering process is to enhance the selection of the barring rate by reducing the effect of the inherent randomness of the distribution of UE accesses and of the RAP. The main contributions of our ACBC scheme are as follows.

1. It effectively operates with minimal information regarding the signaling traffic intensity. In fact, only the number of successful accesses per RAO and the total amount of available resources are needed to accurately set the barring parameters. These values are clearly known by the eNB.

2. It efficiently tolerates the long period between SIB 2 transmissions. As mentioned above, the barring parameters are exclusively broadcast through the SIB 2, whose shortest period is 80 ms, as defined in the specifications [10]. In a typical configuration of the physical RACH (PRACH), RAOs occur once every 5 ms.

Therefore, the period of the SIB 2 is typically 16 times longer than the period of RAOs [10].

3. It successfully configures the barring parameters of the ACB scheme as defined in the RRC specification [10]. Hence, it efficiently relieves congestion given that the ACB scheme only affects the UEs that have not yet begun the RAP.

Our ACBC scheme incorporates the simple and robust least-mean-square (LMS) algorithm to continuously adapt the weights of a filter according to the perceived signaling traffic intensity. Two different configurations of the LMS are considered. The first one is a typical adaptive line enhancer (ALE) configuration, whose purpose is to remove a wideband noise from a narrowband information-bearing signal. The second one is a novel twist on the typical ALE configuration. Preliminary results on the performance analysis of our ACBC scheme with this second configuration can be found in [66].

Initial tests were performed with a different adaptive algorithm, namely the recursive least-squares (RLS), and with finite-duration impulse response (FIR) filters with fixed weights. Nevertheless, the benefits of the latter approaches were lesser when compared to those provided by the LMS algorithm. Results derived from our tests with the RLS algorithm are included in Appendix C.

We assess the performance of our ACBC scheme by means of the idealized scheme with full state information presented by Duan *et al.* [35]; it was also used for benchmark purposes in their work. This scheme cannot be implemented in cellular networks but serves as an upper bound to the performance of ACBC schemes. Results show that a remarkable performance can be obtained by implementing our ACBC scheme with either of the two adaptive filter configurations. For instance, the probability of successfully completing the RAP during periods of congestion can go from a poor 31.3 percent with no implemented ACB scheme to more than 95 percent with our ACBC scheme. In addition, the difference in the access delay of UEs when compared to the benchmark scheme under these conditions is minor. Moreover, the access delay is not affected during periods of no congestion. These characteristics make our ACBC scheme one of the few efficient and practical solutions to congestion caused by mMTC applications in cellular networks.

The rest of the chapter is organized as follows. Section 4.2 presents a review of state-of-the-art solutions to congestion in cellular networks under mMTC scenarios. Next, Section 4.3 presents our ACBC scheme; this includes the adaptive filter algorithms and configurations it incorporates. Section 4.4 describes the test scenarios and the methodology we use to optimize our ACBC scheme. Afterwards, Section 4.6 presents the achievable performance with our ACBC scheme. In this section we use an idealized ACBC scheme for benchmark purposes; we also evaluate the impact of realistic assumptions on the performance of the latter scheme. Finally, we present our conclusions.

4.2 Related work

One of the most promising and widespread approaches to implement an ACBC scheme is to estimate the total number of contending UEs [12, 34, 35, 51, 94, 108]. In most cases, the number of successful and collided preambles is used for this purpose. With this information, an optimal barring rate, typically defined as the barring rate that maximizes the expected number of preambles transmitted by exactly one UE per RAO [108]. Highlights of some of the studies that follow the approach described above are summarized as follows.

Wang *et al.* provide a closed-form approximate solution to the problem of obtaining the optimal barring rate is presented in [108]. Jin *et al.* propose a pseudo-Bayesian ACBC scheme in which the number of idle preambles is used for the estimation [51]. Tavana *et al.* use a Kalman filter to enhance the accuracy of the estimation [94]. While this latter study only considers the first step of the RAP, preamble transmission, the idea of using adaptive filters for this purpose is promising.

Abbas *et al.* proposed an ACBC scheme that considers the use of different barring rates for different groups of UEs based on their delay requirements [12]. As we will observe in Section 4.6, the access delay of UEs under any ACB scheme and during periods of congestion is only suitable for delay tolerant applications, even when the optimal barring parameters are selected. Hence, the potential delay gains of using different barring parameters for different groups of UEs are minimal.

Duan *et al.* propose an ACBC scheme that employs the number of successful and idle preambles, along with the total number of UEs registered previously for the estimation [35]. An extension of this latter approach is also provided in the same paper to dynamically select the number of available preambles allocated to machine-type communications (MTC) UEs.

The performance of the above mentioned ACBC schemes is typically compared with that of idealized solutions that exploit the benefits of having full state information and can make *a priori* decisions [35, 94, 108]. Clearly, these full state information solutions cannot be implemented at the eNBs but provide an upper bound to the performance of the ACB scheme. As it will be seen in Section 4.6, we will adopt this approach and compare the performance of our ACBC scheme with that of the idealized and full state information scheme described by Duan *et al.* [35].

Results presented from most of the studies listed above [35, 94, 108] show that the performance of the proposed ACBC schemes is close to that of the idealized solution. Nevertheless, these ACBC schemes cannot be implemented in 3GPP cellular networks because were developed based on an idealized ACB scheme in which every UE is subject to the ACB scheme, even after the beginning of the RAP and the barring rate is calculated and broadcast by the eNB at each RAO. While the first assumption is a simplification of the ACB scheme that overrides the backoff procedure, it is certainly not possible to broadcast a newly calculated barring rate at each RAO. As mentioned throughout this dissertation, the shortest period of the SIB 2 (where the barring parameters are broadcast) is 80 ms. This is, under a typical PRACH configuration, 16 times longer than the period of RAOs.

Yet another factor that hinders the implementation of the ACBC schemes mentioned above is that, in order to accurately approximate the number of contending UEs, the eNB must be aware of the number of successful preambles and also of at least one of the following: 1) the number of preambles not transmitted by any UE (i.e., idle preambles); or 2) the number of preambles transmitted by more than one UE (i.e., collisions). In a real world implementation, this information may not be available and the reasons for this are manifold. For instance, the eNB may not be able to decode the preambles transmitted by multiple UEs, or some preambles transmitted by exactly one

UE may be lost due to a wireless channel error. mode details on this matter can be found in Chapter 3 and in studies such as Wei *et al.* [109] and de Andrade *et al.* [34].

Lin *et al.* follow a different approach: the use of a state transition diagram for the dynamic activation of the ACB scheme [71]. That is, the state of the system depends on the average number of successful preamble transmissions and the ACB scheme is activated when the system reaches the state of severe congestion. This approach is simpler to implement and analyze as it does not depend on an accurate approximation of the contending UEs. However, the authors do not consider that the number of available uplink grants is limited and is, in a typical configuration, the main bottleneck of the RAP [1]; this will be described in detail in Section 4.4. Furthermore, the performance of the presented ACB scheme is only assessed in terms of the success probability, while other key performance indicators (KPIs) are neglected.

De Andrade *et al.* [34] proposed and evaluated the performance of an ACBC scheme, along with several other access control schemes. The presented schemes consider the delay of each notification mechanism and their performance is assessed in terms of numerous KPIs. Results show that their ACBC scheme leads to the highest success probability (i.e., the probability to successfully complete the RAP) under a highly congested scenario. Nevertheless, the obtained success probability is lower than 0.8. Throughout this chapter, as in previous ones, we assume the target success probability is 0.95.

Finally, Tello-Oquendo *et al.* [97] presented an ACBC scheme that incorporates a reinforcement learning technique. The proposed ACBC scheme may indeed be implemented in cellular networks as it was designed with the restrictions described above. For instance, the shortest period of the SIB 2 was considered. On the other hand, the results obtained with this ACBC scheme were not entirely satisfactory. That is, a sufficiently high success probability can be obtained with this ACBC scheme under a highly congested scenario, but the access delay is more than 25 percent longer when compared to a near-optimal implementation of the fixed ACBC scheme.

Results presented by de Andrade *et al.* [34] and by Tello-Oquendo *et al.* [97] showcase the difficulty of designing ACBC schemes and the impact that the delay of notification mechanisms have on performance.

The work and results presented in the two previous chapter have lead to the development of our novel adaptive ACBC scheme. As will be shown in Section 4.3, our ACBC scheme can be directly implemented in cellular networks as it considers each and every one of the system limitations and can lead to a success probability higher than 95 percent while maintaining a near-optimal access delay.

4.3 Adaptive ACBC scheme

In this section we describe in detail the operation of our novel ACBC scheme. It is important to emphasize that one of its remarkable features is that it strictly adheres to the ACB scheme as defined in the specifications [5, 10]. That is, we provide an efficient method to calculate adequate parameters for the ACB scheme defined in the 3GPP standards. Therefore, it can be directly implemented at the eNBs in 3GPP cellular systems.

The block diagram shown in Fig. 4.1 describes the operation of the RAP with our ACBC scheme. From Fig. 4.1 two main blocks can be clearly identified: RA and ACBC. Depicted in the upper part of Fig. 4.1 is the RA, which comprises the barring checks and the RAP. These were thoroughly described in Chapter 2.3. Clearly, the RA can be initiated at any RAO. Therefore, the discrete time index i of the variables involved in the RAP stands for the epoch number, where the epoch duration is one RAO.

Fig. 4.1 introduces $x(i)$, defined as the number of UEs that attempt to switch from idle to connected mode for the first time at the ith RAO (i.e., UE arrivals). That is, $x(i)$ is outcome of a single experiment for random variable (RV) X_i. Before initiating the RAP, these UEs must perform a barring check. The UEs that fail the barring check consider their access as barred and must wait for a random time, calculated as $t_w = (0.7 + 0.6\ \mathrm{U}[0,1))\, t_{\mathrm{acb}}(j)$. This process is performed until a barring check is successful, when the UE is allowed to initiate the RAP

Next, please recall $n(i,k), \quad k \in \{1,2,\ldots,k_{\max}\}$ is defined as the number of UEs that transmit the kth preamble at the ith RAO. Then, $n(i,1)$ denotes the number of UEs whose first preamble is transmitted at the ith RAO. That is, the UEs whose barring

check at the ith RAO is successful; these UEs initiate the RAP immediately. Finally, we define $g(i)$ as the number of UEs that have successfully transmitted a preamble at the ith RAO and that will receive an uplink grant within the ith (i.e. next) RA response (RAR) window. Strictly speaking, $g(i)$ is the observed value of the stochastic process $\{N_{G,i}\}_{i\in\mathbb{N}}$ at time index i, defined in (3.30) on page 59. It was observed in Chapter 3 that the probability that a UE successfully completes the RAP, given an uplink grant was received, closely approaches one. Hence, $g(i)$ is a tight upper bound to the number of successful accesses at the ith RAO that can be calculated immediately after the eNB decodes the preambles 2 ms after transmission and is a close approximation to the number of UEs that successfully complete the RAP. On the other hand, the eNB would have to wait until the end of the RAP (e.g., at least 15 ms given $t_{\text{rao}} = 5$ ms) to obtain this information.

In the ACBC block, depicted in the lower part of Fig. 4.1, the eNB calculates the barring parameters that will be broadcast through the jth SIB 2: mean barring time $t_{\text{acb}}(j)$ and barring rate $p_{\text{acb}}(j)$. The SIB 2 is broadcast once every t_{si} RAOs, hence, these parameters are adapted according to the perceived signaling traffic intensity throughout this period. As such, the discrete time index j stands for the epoch number when the epoch duration is t_{si} RAOs. Consequently, the ACBC block operates at a time scale that is t_{si} times greater than that of the RA block. Specifically, the jth SIB 2 is broadcast at the $i = (jt_{\text{si}})$th RAO. As such, $p_{\text{acb}}(j)$ and $t_{\text{acb}}(j)$ remain constant from the $(jt_{\text{si}} + 1)$th until the $(\lfloor j + 1 \rfloor t_{\text{si}})$th RAO. We now describe in detail the process to calculate the barring parameters.

The eNB calculates the ratio of utilized to available resources immediately before the jth SIB 2 transmission. For this, let $n(i)$ be the number of contending UEs (i.e., total number of preamble transmissions) at the ith RAO. Please recall the theoretical capacity of the RAP $\mathrm{C}(r, g)$ is defined as the maximum expected number of uplink grants transmitted within a RAR window that can be obtained for a given number of available preambles r, uplink grants g, and for any $n(i) \in \mathbb{R}_{\geq 0}$ (Definition 2.3.2 on page 22, Chapter 2). This capacity is achieved when the number of contending UEs is $n(i) = \lceil \log(r/(r-1)) \rceil^{-1}$ and is calculated in (2.8) as the minimum between the PRACH and physical downlink control channel (PDCCH) capacities.

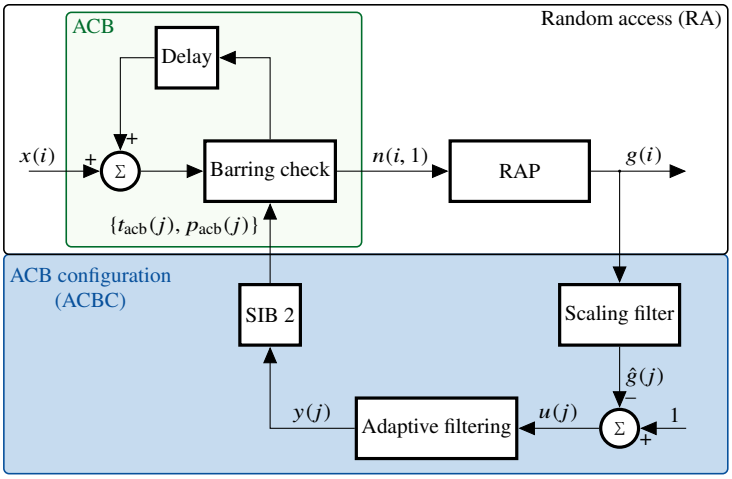


Figure 4.1: Block diagram of the RAP with our novel ACBC scheme. The random access is performed at each RAO, whereas the ACBC can only be performed once every t_{si} RAOs.

With this information, the ratio of utilized to available resources is calculated in the scaling filter block shown in Fig. 4.1 immediately before the jth SIB 2 transmission (i.e., at the (jt_{si})th RAO) as

$$\hat{g}\,(j) = \frac{1}{t_{\mathrm{si}}\ \mathrm{C}\left(r, n_{\mathrm{ug}}\right)} \sum_{i=(j-1)t_{\mathrm{si}}+1}^{jt_{\mathrm{si}}} g\,(i)\,. \tag{4.1}$$

Then, let $u(j)$ be the ratio of idle to available resources for the jth SIB 2 broadcast interval. It is easily calculated as

$$u(j) = 1 - \hat{g}\,(j) \tag{4.2}$$

and serves as the input to the adaptive filtering process. The filter output $y(j)$ is used to calculate the barring rate for the jth SIB 2 broadcast interval

$$p_{\mathrm{acb}}(j) = \min\{y(j), 1\}\,. \tag{4.3}$$

Hence, the $p_{\mathrm{acb}}(j)$ with the ratio of idle to available resources. This increases the

probability of delaying the beginning of the RAP when most of the resources have been utilized.

Finally, we propose the dynamic selection of the mean barring time $t_{\text{acb}}(j)$ as a function of $p_{\text{acb}}(j)$. For this, let $t_{\max}$ be the longest mean barring time that can be broadcast by the eNB. Hereafter we refer to $t_{\max}$ simply as the barring indicator; it is fixed and selected empirically by the network administrator. Then, the mean barring time is calculated as

$$t_{\text{acb}}(j) = (1 - p_{\text{acb}}(j))^{\omega}\ t_{\max}; \tag{4.4}$$

where exponent $\omega \in \mathbb{R}_{\geq 0}$. The impact of parameter ω on the performance of our ACBC scheme is discussed in Section 4.6. We now proceed to describe the selected adaptive algorithm and the two different configurations that were implemented in our ACBC scheme.

4.3.1 Adaptive filter algorithm configurations

The LMS is an adaptive filter algorithm that is widely used because of its simplicity and numerical robustness [45]. Concretely, the complexity of the LMS algorithm is $O(\ell)$, where ℓ is the filter length. That is, its complexity scales linearly with the filter length as $2\ell + 1$ multiplications and $2\ell + 1$ additions are performed per iteration (in our case, per ACBC process) [45]. Since the eNBs possess great computational power, they can easily implement the LMS algorithm.

The block diagram of the LMS adaptive filter algorithm is shown in Fig. 4.2. A buffer has been incorporated to clearly illustrate that the ratio of idle resources during the last ℓ SIB 2 intervals

$$\boldsymbol{u}(j) = [u(j), u(j-1), \ldots, u(j-\ell+1)] \tag{4.5}$$

serves as the input to the algorithm. In other words, a single value of $u(j)$ is the input to the buffer (as indicated by the thin arrow in Fig. 4.2) and the output of the buffer is a vector (as indicated by the thick arrows).

Fig. 4.2 also shows that the LMS algorithm consists of two processes: the filtering and the adaptive process, which result in a feedback loop. In the filtering process, the

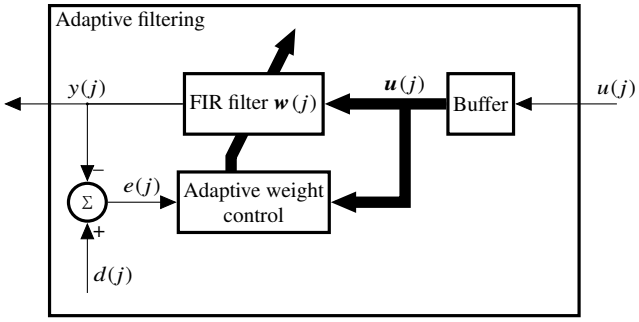


Figure 4.2: Block diagram of the LMS adaptive filter algorithm.

output of a finite-duration impulse response (FIR) filter $y(j)$ is computed from $\boldsymbol{u}(j)$.

In the adaptive process, the output $y(j)$ is compared to the desired response $d(j)$ to obtain the *a priori* error $e(j)$. Then, $e(j)$ serves as an input to the adaptive weight control mechanism. The latter is in charge of adapting the weights of the FIR filter

$$\boldsymbol{w}(j) = [w_0(j), w_1(j), \ldots, w_{\ell-1}(j)] \tag{4.6}$$

automatically based on $e(j)$ and on the adaptation step size μ. The LMS adaptive filter algorithm is summarized in Algorithm 2.

It is important to mention that μ determines the so-called energy constraint or rate of adjustment α, which links the *a priori* error $e(j)$ with the *a posteriori* error $\varepsilon(j)$ as follows [102, Chapter 5.3].

$$\begin{aligned} \varepsilon(j) &= d(j) - \boldsymbol{u}^{\mathsf{T}}(j)\, \boldsymbol{w}(j+1) \\ &= d(j) - \boldsymbol{u}^{\mathsf{T}}(j)\, (\boldsymbol{w}(j) + \mu\, \boldsymbol{u}(j)\, e(j)) \\ &= (1 - \alpha(j))\, e(j); \end{aligned} \tag{4.10}$$

where $\alpha(j) = \mu \|\boldsymbol{u}(j)\|^2$ is the energy constraint at time index j and $\|\cdot\|$ is the Euclidean norm operator. As such, parameter α determines the rate at which $\boldsymbol{w}(j)$ is adjusted, based on $\boldsymbol{u}(j)$.

For the LMS algorithm to be stable, the value of μ must satisfy [45, 102, 111]

$$|1 - \alpha(j)| \leq 1 \qquad \forall j, \tag{4.11}$$

Algorithm 2 LMS adaptive filter algorithm.

Require: the filter length ℓ

Require: the adaptation step size μ

1: Initialize the vector of filter coefficients $\boldsymbol{w}\,(0)$ and the input vector $\boldsymbol{u}\,(0)$ as

$$w_m\,(0) = u\,(-m) = 0, \qquad m \in \{0, 1, \ldots, \ell - 1\} \tag{4.7}$$

2: **for all** $j = 1, 2, \ldots$ **do**

3: Select the desired response $d(j)$

4: Filtering process:

$$y(j) = \boldsymbol{w}^{\mathsf{T}}(j)\boldsymbol{u}(j) \tag{4.8}$$

5: Adaptive process:

$$e(j) = d(j) - y(j) \tag{4.9a}$$

$$\boldsymbol{w}\,(j+1) = \boldsymbol{w}(j) + \mu e(j)\boldsymbol{u}(j) \tag{4.9b}$$

6: **end for**

which gives

$$0 < \alpha_{\max} \leq 2; \tag{4.12}$$

where $\alpha_{\max} = \mu \max\left\{\|\boldsymbol{u}(j)\|^2\right\}$ for all j.

Please observe that, in our ACBC scheme, $\alpha_{\max}$ is achieved when no UE arrivals occur during ℓ consecutive RAOs. In such case, all the resources during are idle, which gives $u(j-m) = 1$ for $m \in \{0, 1, \ldots, \ell - 1\}$; hence, we have $\alpha_{\max} = \mu\ell$. Building on this, the possible values of μ are bounded by the inequality

$$0 < \mu \leq \frac{2}{\ell}. \tag{4.13}$$

One of the most typical applications of the LMS adaptive algorithm is that of an ALE. An ALE is a system that may be used to detect a sinusoidal or narrowband

information-bearing signal buried in a wideband noise background [45, Chapter 6]. In our ACBC scheme, sudden variations of $u(j)$ represent the wideband noise, in which the narrowband information signal is buried. In other words, $u(j)$ is affected by the randomness of both, the distribution of UE arrivals and of the RAP. Hence, the filter weights are automatically adjusted by the LMS algorithm to suppress the sudden variations of $u(j)$.

In this study we propose and evaluate the performance of two different configurations of the LMS ALE. The first one is the typical ALE configuration and the second one, is a novel twist in the ALE configuration that causes the LMS algorithm to "pull" towards a desired output which is selected empirically. Hereafter we refer to the latter as the "pulling" ALE (PALE) configuration. These two configurations are now described in detail.

ALE: This a typical ALE configuration, in which the desired response (primary input) is the ratio of idle to available resources calculated at the jth SIB 2 broadcast interval $d(j) = u(j)$, while the (reference) input is a delayed version of the latter. That is, the input of the algorithm is $u(j-\Delta)$, where Δ is the *decorrelation delay*. Therefore, the input vector is given as

$$\boldsymbol{u}(j-\Delta) = \left[u\left(j-\Delta\right), u\left(j-\Delta-1\right), \ldots, u\left(j-\Delta-\ell+1\right)\right]. \tag{4.14}$$

By implementing the ALE configuration, the filter weights are automatically adjusted to minimize the error between $u(j)$ and its past values $u(j-\Delta-m)$ for $m \in \{0, 1, \ldots, \ell-1\}$. As a consequence, sudden variations are suppressed from $y(j)$. To implement this configuration, it is sufficient to substitute $d(j)$ with $u(j)$, and $\boldsymbol{u}(j)$ with $\boldsymbol{u}(j-\Delta)$ in equations (4.8), (4.9a), and (4.9b) of Algorithm 2. Fig. 4.3 shows the block diagram of the ALE with the LMS adaptive algorithm.

A consideration of importance is to set Δ to a sufficiently large value, so the noise in $u(j)$ is not correlated with that in $u(j-\Delta)$. We have observed that, since t_{si} is large when compared to t_{rao}, it is sufficient to set $\Delta = 1$.

PALE: This is a new twist on the typical ALE configuration, in which the desired response $d(j)$ is set to be a constant selected empirically. On the other hand, the

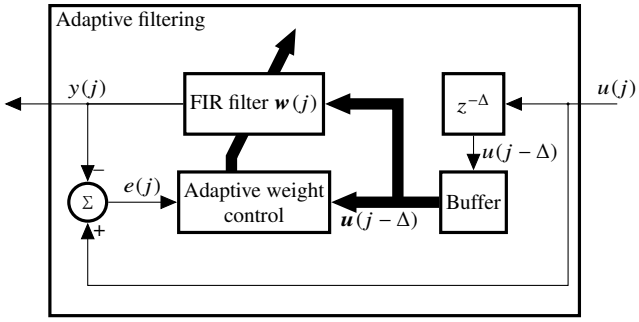


Figure 4.3: Block diagram of the ALE with the LMS adaptive algorithm.

(reference) input is simply $u(j)$. Clearly, no correlation exists between the constant $d(j)$ and the variations in $u(j-\Delta)$ for all $\Delta \in \mathbb{N}$. Therefore we can set $\Delta = 0$.

By implementing the PALE configuration, the filter weights are automatically adjusted to minimize the error between $d(j)$ and $u(j)$, and to suppress the sudden variations of the latter. As a result, $p_{\text{acb}}(j)$ is "pulled" towards $d(j)$. Building on this, we suggest to set $d(j) = 1$. That is, equal to the maximum value of $u(j)$, which is only obtained when all the resources during the SIB 2 broadcast interval are idle. As it will be seen in Section 4.6, setting $d(j) = 1$ minimizes the delay of UEs during intervals of low signaling traffic intensity.

4.4 Test scenarios, tools, and methodology

Access control schemes must provide an adequate performance under several traffic conditions and network configurations. Hence, we evaluate the benefits of our ACBC scheme under the two mMTC traffic models suggested by the 3GPP for the performance evaluation of the RAP [1]. These models were introduced in Table 2.1 on page 23. As a matter of summary, we denote a period of 60 seconds in which $n = 30\,000$ UE arrivals occur according to a uniform distribution as the traffic model (TM) 1; this represents a low signaling traffic load interval. Moreover, we denote a period of 10 seconds in which $n = 30\,000$ UE arrivals occur according to a Beta$(3, 4)$ as the TM 2;

this represents a high signaling traffic load interval. The TM 2 is the one that has attracted the most attention from the research community [30, 71, 109] and is the one we have focused in previous chapters. For instance, we have observed in Chapter 2 that only 31.31 percent of the UEs successfully complete the RAP when a typical configuration is selected.

Throughout the following section, we adopt the most typical configuration of the PRACH and PDCCH channels, presented in Table 3.1, but also incorporate the case in which the number of available preambles $r = 30$, besides the traditional $r = 54$. The reason to incorporate the former value is that fewer preambles are available in narrowband Internet of Things (NB-IoT) (i.e., 48) when compared to traditional LTE Advanced (LTE-A) (i.e., 64). As the eNB commonly reserves some preambles for high priority UEs in LTE-A and for higher coverage enhancement (CE) levels in NB-IoT, selecting $r = 30$ for CE level zero in NB-IoT allows for the reservation of the remaining 18 preambles for UEs with a higher CE level. As described in Section 2.3 on page 12, the number of UEs in CE level zero is expected to be far greater than those in CE levels one and two. Hence, these contribute the most to the signaling traffic intensity and congestion.

The two possible values of r are quantitatively different from the perspective of our ACBC scheme. To showcase this difference, Fig. 4.4 shows the expected number of assigned uplink grants at the ith RAR $\mathbb{E}[N_G]$ window as a function of $n(i)$ when $r \in \{30, 54\}$ preambles and $g = 15$ uplink grants are available; the capacity of the RAP for these two combinations of r and g is also included. These results were obtained by means of the analytical model presented in Chapter 3 [60, 69] and by (2.8) on page 22, respectively. As it can be seen, $\mathbb{E}[N_G]$ is a concave function whose global maximum is exactly at $n^*(i) = \lceil \log(r/\lceil r-1 \rceil) \rceil^{-1}$, same as in (2.3). In it worth mentioning that the curve for $r = 30$ in Fig. 4.4 highly resembles that in Fig. 2.2 on page 19 for the same r, where the expected number of successful preambles is shown. On the other hand, there is a noticeable difference between the curves for $r = 54$ in these two figures. Furthermore, the rate of change when $r = 30$ is higher than when $r = 54$, especially as $n(i) \to \lceil \log(r/(r-1)) \rceil^{-1}$. The main reason for this is that $C(30, 15)$ (i.e., the theoretical capacity of the RAP for $r = 30$) is limited by r. That is, $C(30) = 11.23 < 15$, hence $C(30, 15) = 11.23$. Conversely, $C(54) = 20.05 > 15$ is

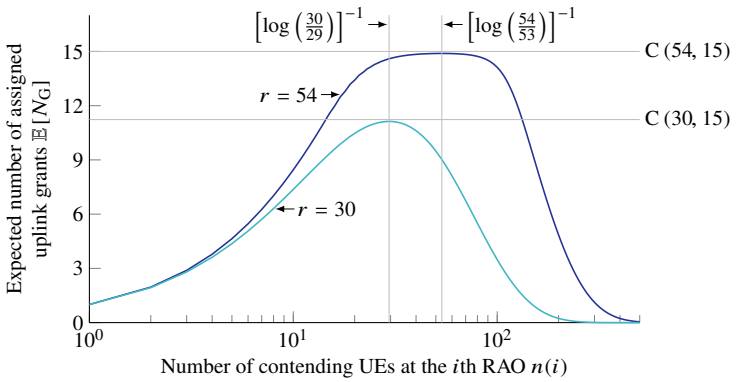


Figure 4.4: Expected number of assigned uplink grants at the ith RAR window given r available preambles, $g = 15$, and $n(i)$ contending UEs. The x-axis is shown in logarithmic scale.

clearly limited by the g, hence $\mathrm{C}(54, 15) = 15$. As will be confirmed throughout the following section, this in turn makes the adequate configuration of our ACBC scheme more challenging for $r = 30$ than for $r = 54$.

For the results presented in the following section we select $t_{\mathrm{si}} = 16$ RAOs as the period of the SIB 2. This value is the result of selecting the most typical period of RAOs $t_{\mathrm{rao}} = 5$ ms and the minimum period for SIB transmissions, 80 ms. Please recall the SIB 2 carries the barring parameters, hence, we have selected the maximum frequency of update for these parameters.

Results were obtained by incorporating the adaptive algorithm configurations presented in Section 4.3 to the simulator used to obtain the results presented in previous chapters. In each simulation, the adaptive algorithm is initialized as described previously and the filter weights are stabilized. Then, $n = 30\,000$ UE arrivals are scheduled within the distribution period t_{dist}, which begins at $i = 0$. The jth SIB 2 is broadcast at the $(jt_{\mathrm{si}} + i_r)$th RAO, where $i_r = \mathrm{U}[0, t_{\mathrm{si}} - 1]$ is a discrete random time shift. A simulation run ends when every UE has terminated the RAP. As in previous chapters, the number of simulation runs is set to the smallest number that ensures that all the cumulative KPIs obtained up to the last simulation differ from those obtained up to the

previous simulation by less than 0.01 percent.

4.4.1 Performance metrics and methodology

The performance of the RA with our adaptive ACBC scheme is assessed in terms of the following KPIs.

1. Success probability P_s, defined as the probability to successfully complete the RAP within the maximum number of preamble transmissions.

2. Access delay D, defined as the time elapsed between the first access attempt (barring check or preamble transmission) of a UE and the successful completion of its RAP. It is assessed in terms of the 95th percentile D_{95} given in seconds. That is, the delay of 95 percent of the UEs that successfully complete the RAP is D_{95} or less. The performance under the TM 1 is assessed in terms of the increase in delay due to the implementation of an ACBC scheme given as $\Delta D_{95} = D_{95} - D^*_{95}$, where D^*_{95} is the 95th percentile of access delay obtained with no ACB scheme for the selected configuration.

3. Number of preamble transmissions performed by the UEs that successfully complete the RAP K. It is assessed in terms of its expected value $\mathbb{E}[K]$.

The methodology for our study is as follows. We first find an adequate value of parameter μ. For this, we observe the behavior of our ACBC scheme under the TM 1 for different values of μ in the range defined by (4.13). An adequate value of μ is selected empirically based on the response from the adaptive filter. Specifically, we aim to suppress the sudden variations of $u(j)$ while achieving the fastest possible convergence toward its expected value $\mathbb{E}[u(j)]$. It is important to emphasize that similar trial and error approaches to select μ are the most common in the practice [110]. The reason for this is that μ depends on several factors that are application-specific and may not be known. For example, adequate values of μ highly depend on the variability of the input. Consequently, the selection of an optimal value is oftentimes impossible.

Then, we continue to find the "optimal" configuration of our ACBC scheme. It is defined as the combination of the barring indicator $t_{\max}$, filter length ℓ, and exponent

ω that leads to the minimum D_{95} under both traffic models given $P_s \geq 0.95$ under the TM 2 for a given r. We denote the optimal values of these parameters as $t^*_{\max}$, ℓ^*, and ω^*, respectively.

Duan *et al.* [35] proposed an idealized full state information scheme that is used as a benchmark for their ACBC scheme; the latter is called D-ACB. As described by Duan *et al.* [35], the benchmark scheme has full state information on the number of contending UEs at each RAO, hence, it can select the optimal barring rate accordingly. On the other hand, their D-ACB scheme estimated the number of contending UEs based on the number of successful and idle preambles, but also on the number of previously registered UEs.

It is important to observe that the every ACBC scheme presented by Duan *et al.* [35] was designed for an idealized ACB scheme. That is, they assume the barring parameters are calculated and transmitted at each RAO and also that every UE is subject to the ACB scheme even after initiating the RAP. This is not the behavior of the ACB scheme as defined in the protocol specifications [10].

We have extended the original benchmark scheme proposed by Duan *et al.* [35] to cope with the periodicity of the SIB 2 t_{si}. As such, the optimal barring rate is calculated as

$$p^*_{\text{acb}}(j) = \min\left\{1, \frac{r}{n'(j)}\right\} \tag{4.15}$$

where

$$n'(j) = \frac{1}{t_{\text{si}}} \sum_{i=(j-1)t_{\text{si}}+1}^{jt_{\text{si}}} n(i)\,; \tag{4.16}$$

please recall that $n(i)$ is the number of contending UEs at the ith RAO.

Also, please observe that (4.15) is exactly as defined by Duan *et al.* [35] for $t_{\text{si}} = 1$, and we simply introduce $n'(j)$ to obtain the average optimal barring rate for any $t_{\text{si}} \geq 1$ RAO. Hereafter we refer to this extended scheme simply as the idealized full state information (IFI) scheme; it is used to assess the performance of our ACBC scheme. The barring time $t_{\text{acb}}(j)$ at each barring check under both schemes is deterministic of one RAO [35].

In the following section, we present relevant results derived from the performance analysis of both, the ALE and PALE configurations, along with their optimal parameter configurations.

4.5 Results and discussion

In this section we present and discuss relevant results obtained from the performance evaluation of the RAP with our ACBC scheme. As a starting point, we find an adequate value for parameter μ. Then, we compare the performance of our ACBC scheme with that of: 1) our ACBC scheme with no filtering process; 2) a static ACBC scheme with fixed $p^*_{\text{acb}}(j)$ and $t^*_{\text{acb}}(j)$ as in previous chapters; and 3) the IFI scheme. The optimal configuration of each of these schemes is assumed. Next, showcase the robustness of our ACBC scheme by evaluating the impact that deviations from the optimal configuration have on performance. Finally, we discuss the impact of realistic assumptions on the performance of the IFI scheme.

We investigate the impact of μ on the response of the adaptive algorithm by observing its behavior under the TM 1. For this, Fig. 4.5 shows the response of the algorithm during the first 100 SIB 2 transmissions with the ALE configuration for $\mu \in \{2/\ell, 1/(25\ell), 1/(50\ell)\}$. Results from a single simulation run are shown to showcase the impact of μ; we have confirmed that these results represent the common behavior of the adaptive algorithm. Typical values $\ell = 32$ and $r = 54$ have been selected and UEs ignore the ACB scheme (e.g., were assigned to high priority ACs). That is, at this point we are only interested in observing the difference between the calculated $u(j)$ and $p_{\text{acb}}(j)$, not in their effect in the UE arrivals.

In particular, we are set to find a setting for μ that successfully reduces the variations of $u(j)$ with the fastest possible convergence toward $\mathbb{E}\left[u(j)\right]$. Under the traffic model 1, $n = 30\,000$ UE accesses are uniformly distributed within 60 s. Next, please recall $\{X_i\}_{i\in\mathbb{N}}$ is the stochastic process that defines the number of UE accesses at each RAO within the distribution period. Hence, $x(i)$ is the outcome of a single experiment for RV X_i. Given $t_{\text{rao}} = 5$ ms, we have $\mathbb{E}\left[X_i\right] = n/12\,000 = 2.5$ s.t. $\{i \in \mathbb{N} \mid i < i_{\text{dist}}\}$, where i_{dist} is the last RAO in the distribution period. From there, the following approximation

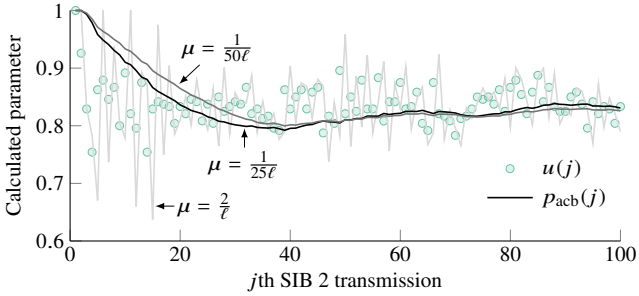


Figure 4.5: Ratio of idle to available resources $u(j)$ and barring rate $p_{\text{acb}}(j)$ calculated at the jth SIB 2 for $\mu \in \{2/\ell, 1/(25\ell), 1/(50\ell)\}$; UEs ignore the ACB scheme.

can be obtained by substituting $g(i)$ with $\mathbb{E}[X_i]$ in (4.1) and (4.2).

$$\mathbb{E}[u(j)] \approx 1 - \frac{1}{t_{\text{si}}\,\mathrm{C}(r,g)} \sum_{i=(j-1)t_{\text{si}}+1}^{jt_{\text{si}}} \mathbb{E}[X_i] = 1 - \frac{\mathbb{E}[X_i]}{\mathrm{C}(r,g)} \tag{4.17}$$

which gives $\mathbb{E}[u(j)] \approx 5/6$ for $r = 54$. This value has been confirmed by simulation and by the analytical model presented in Chapter 3 [69].

It can be seen in Fig. 4.5 that the maximum possible value of $\mu = 2/\ell$ does not provide the desired response because the variations of $p_{\text{acb}}(j)$ are even greater than that of $u(j)$. On the other hand, the LMS algorithm behaves as a low-pass filter with a sharp cutoff bandwidth that successfully suppresses the sudden variations of $u(j)$ when lower values of μ are selected. However, it can also be observed in Fig. 4.5 that $\mu = 1/(50\ell)$ induces a slightly higher delay than $\mu = 1/(25\ell)$. That is, the curve for $\mu = 1/(50\ell)$ converges more slowly toward $\mathbb{E}[u(j)]$ than the curve for $\mu = 1/(25\ell)$ and the variations of both are comparable. Hence, $\mu = 1/(25\ell)$ is used throughout the remainder of the paper. The interested reader is referred to [45, Chapter 6] for more details on the impact of μ in the response of the LMS algorithm.

Table 4.1: Optimal configuration of the different ACBC schemes.

ACBC scheme	Parameter	Optimal value	
		$r = 30$	$r = 54$
ALE	ω	3	3
	ℓ	32	32
	$t_{\max}$	3.8	0.3
PALE	ℓ	16	32
	$t_{\text{acb}}(j) = t_{\max}$	4.2	0.6
No filtering	ω	0	2
	$t_{\max}$	5.4	3.3
Static	$p_{\text{acb}}(j)$	0.11	0.31
	$t_{\text{acb}}(j)$	1.2	1.75

4.5.1 Performance of ACBC schemes with the optimal configuration

We begin our performance analysis by presenting the optimal configuration of the selected ACBC schemes given $t_{\text{si}} = 16$ RAOs in Table 4.1. As mentioned above, the optimal configuration of each ACBC scheme is defined as the configuration that leads to the shortest D_{95} under both traffic models given $P_s \geq 0.95$ under the TM 2. To find the optimal configuration of our ACBC scheme, we have evaluated the performance with $\omega \in \mathbb{N}$, $\ell \in \{1, 2, 4, \ldots, 128\}$, and $t_{\max} \in \{0.1, 0.2, \ldots, 10\}$ s for each $r \in \{30, 54\}$. We have observed that the optimal value of the mean barring time for the PALE configuration is simply $p^*_{\text{acb}}(j) = t^*_{\max}$.

The KPIs obtained under both traffic models with the optimal configuration of each of the selected ACBC schemes are shown in Table 4.2. KPIs obtained with no implemented ACB scheme have been included as a reference. The success probability P_s under the TM 1 has been omitted because it is equal to one for all cases.

It is important to emphasize that the IFI scheme cannot be implemented in 3GPP cellular networks. As a consequence, the performance reported in Table 4.2 for the IFI scheme is not achievable in practice. However, it provides an upper bound for

Table 4.2: KPIs obtained with the optimal configuration of the selected ACBC schemes and with no ACB scheme under the TM 2.

	Success probability		95th percentile of access delay (s)		Expected number of preamble transmissions	
ACBC scheme	$r = 30$	$r = 54$	$r = 30$	$r = 54$	$r = 30$	$r = 54$
ALE	0.951	0.965	14.450	6.807	2.438	2.584
PALE	0.968	0.979	14.425	7.286	2.557	2.485
No filtering	0.997	0.967	21.440	10.839	2.065	2.189
Static	0.951	0.950	30.927	13.584	2.348	2.635
IFI	0.988	0.971	11.491	5.468	3.123	3.392
No ACB	0.115	0.313	0.175	0.182	3.157	3.452

Table 4.3: KPIs obtained with the optimal configuration of the selected ACBC schemes and with no ACB scheme under the TM 1.

	95th percentile of access delay (s)		Expected number of preamble transmissions	
ACBC scheme	$r = 30$	$r = 54$	$r = 30$	$r = 54$
ALE	0.110	0.057	1.576	1.500
PALE	0.065	0.059	1.575	1.500
No filtering	6.984	0.165	1.567	1.500
Static	30.349	13.379	1.548	1.494
IFI	0.060	0.055	1.576	1.500
No ACB	0.060	0.055	1.575	1.500

the performance of the ACB scheme. A detailed study on the impact of realistic assumptions on the performance of the IFI scheme is presented on page 114.

Table 4.2 reveals that $P_s \geq 0.95$ can be obtained with any of the selected ACBC schemes under the TM 2 for both $r \in \{30, 54\}$. As it can be seen, the D_{95} obtained

with our ACBC scheme with the ALE and PALE configurations is up to 50 percent shorter than with no filtering process. This showcases the benefits of incorporating an adaptive filter. Moreover, the D_{95} obtained with our ACBC under the TM 2 is, in the worst case, around 28 percent longer than the one obtained with the IFI scheme. This difference is significant, but is important to emphasize that our ACBC scheme can be implemented in the eNBs in its current form.

Yet another interesting aspect is that the D_{95} obtained with our ACBC scheme under the TM 2 is around 48 percent shorter than the one obtained with the static ACBC despite the long period $t_{\text{si}} = 16$. However, the achieved D_{95} with any of the ACBC schemes under the TM 2 is in the order of a few seconds; such long delay is only suitable for delay-tolerant applications.

Needless to say, the optimal performance under the TM 1 is obtained with no ACB scheme, but also with the IFI scheme. That is, the effect of the deterministic barring time of one RAO combined with a sufficiently high $p^*_{\text{acb}}(j)$ is not observable in the selected KPIs. On the other hand, the longest D_{95} under the TM 2 is obtained with the static ACBC and a similar D_{95} is obtained under the TM 1. Naturally, the static ACBC is not an efficient solution to congestion.

Table 4.2 also shows that the D_{95} obtained with our ACBC scheme under the TM 1 is less than seven percent higher than the minimum, achieved with no ACB scheme. The only exception occurs with the ALE configuration for $r = 30$. In this case, setting $\omega = 3$ is not sufficient to achieve a lower D_{95} and, as it will be discussed later in this section, selecting $\omega \geq 4$ sharply increases t^*_{max} to the point that there is no $t_{\text{max}} \leq 10$ s that leads to $P_s \geq 0.95$ under the TM 2. Moreover, the effect of increasing ω on t^*_{max} is magnified if no filtering process is incorporated to our ACBC scheme. Concretely, no combination of t_{max} and $\omega > 0$ given $r = 30$ exists for which $P_s \geq 0.95$ and selecting $\omega = 0$ results in an excessively long access delay under the TM 1.

We have also evaluated the performance of our ACBC scheme with the optimal configuration shown in Table 4.1 under congestion scenarios comparable to the TM 2. For instance, when $n = 30\,000$ UE arrivals follow a Beta $(4, 4)$ distribution over $t_{\text{dist}} = 10$ s. The peak in the average number of UE arrivals is around five percent higher for Beta $(4, 4)$ than for Beta $(3, 4)$. The performance of our ACBC scheme under

this traffic model is comparable to that under the TM 2 (see Table 4.2) as $P_s \geq 0.93$ is achieved with both the ALE and PALE configurations for $r \in \{30, 54\}$. Furthermore, the difference in D_{95} between these two traffic models is less than one percent.

Now we proceed to compare the behavior of the ALE and PALE configurations. For this, Fig. 4.6 shows the ratio of idle to available resources $u(j)$ and $p_{\text{acb}}(j)$ with the optimal ALE and PALE configurations given $r = 54$. A similar behavior was observed for $r = 30$, so these results have been omitted.

It is important to point out that the first 12 000 RAOs after the beginning of the distribution period are shown in Fig. 4.6a and Fig. 4.6b as $t_{\text{si}} = 16$ RAOs. On the other hand, the first 3200 RAOs are shown in Fig. 4.6c and Fig. 4.6d. Again, results from a single simulation run are shown and we have confirmed that these correspond to the common behavior of our ACBC scheme.

We can clearly observe in Fig. 4.6 that the filtering process smooths out the sudden variations (noise) of $u(j)$. The result is a much more stable and accurate selection of $p_{\text{acb}}(j)$. Also it can be seen that the calculated $u(j)$ with the ALE configuration under the TM 1 (see Fig. 4.6a) is similar to the one calculated with the PALE configuration (see Fig. 4.6b) despite the fact that for the former $p_{\text{acb}}(j) < 1$ for all j. This is caused by the selection of $\omega^* = 3$ and $t^*_{\text{max}} = 0.3$ s, which results in $t_{\text{acb}}(j) \approx 1 \cdot 10^{-3}$ s for all j, which is negligible.

On the other hand, the "pulling" effect of the PALE configuration can be clearly observed in Fig. 4.6b and in Fig. 4.6d. That is, $p_{\text{acb}}(j) > u(j)$ for most j with the PALE configuration under the TM 2, and for every j under the TM 1. This effect is emphasized by the red arrows, which indicate the difference in amplitude between $u(j)$ and $p_{\text{acb}}(j)$. For instance, Fig. 4.6c clearly shows that $u(20) < p_{\text{acb}}(20)$, while $u(160) > p_{\text{acb}}(160)$ with the ALE configuration. This is caused by the delay in the response of the algorithm. On the other hand, Fig. 4.6d shows that $u(j) < p_{\text{acb}}(j)$ for both $j \in \{20, 160\}$ with the PALE configuration. Although the difference between $u(160)$ and $p_{\text{acb}}(160)$ is barely noticeable.

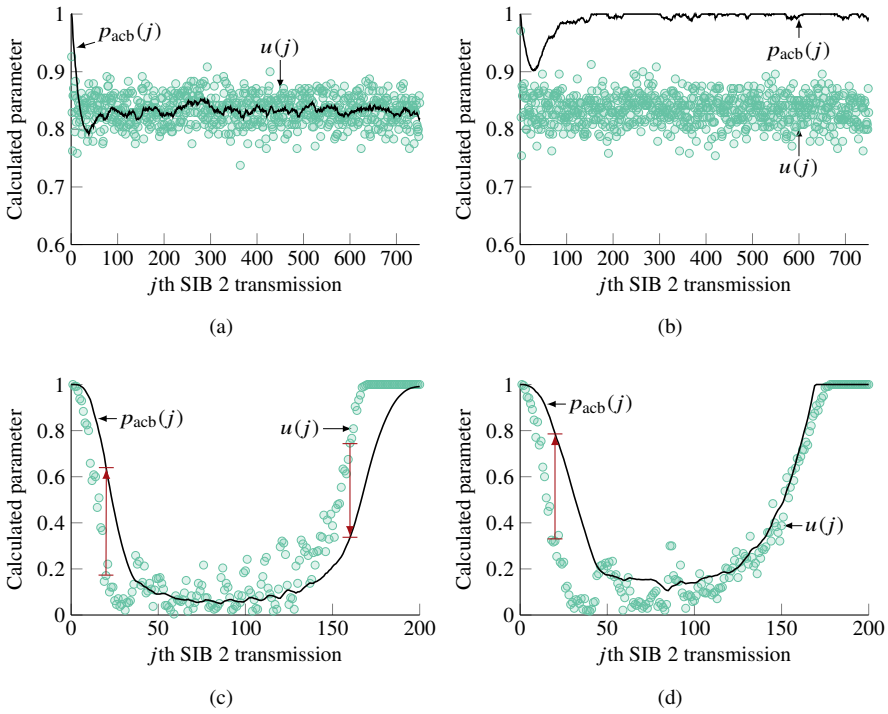


Figure 4.6: Ratio of idle to available resources $u(j)$ and barring rate $p_{acb}(j)$ calculated at the jth SIB 2 for a single simulation run and $r = 54$ for the: (a) ALE, TM 1; (b) PALE, TM 1; (c) ALE, TM 2; and (d) PALE TM 2.

4.5.2 Robustness of the proposed ACBC scheme

In this subsection we showcase the robustness of our ACBC scheme by showing the impact that deviations from the optimal ALE and PALE configurations have on performance.

We first investigate the impact of deviations from the optimal value of ω on the performance of the ALE configuration. For this, the obtained ΔD_{95} (under the TM 1) is shown in Fig. 4.7a and the obtained D_{95} under the TM 2 is shown in Fig. 4.7b for $\omega \in \{0, 1, \ldots, 7\}$ and $r \in \{30, 54\}$, given t^*_{max}. Fig. 4.7 only shows plots corresponding

to $r = 30$ when $\omega \leq 3$ because there is no $t^*_{\max} \leq 10$ s for $\omega > 4$. That is, there is no $t_{\max} \leq 10$ s that leads to $P_s \geq 0.95$ for $\omega \geq 4$ when $r = 30$. This same occurs for $\omega \geq 8$ when $r = 54$.

Fig. 4.7a also shows that high values of ω sharply decrease ΔD_{95} but, as described above, excessively high values of ω may greatly increase $t^*_{\max}$. Building on this, ω must be carefully selected to reduce the access delay under the TM 1, but also to achieve an adequate response under the TM 2, especially if $r = 30$.

Next, we evaluate the impact on performance of deviations from ℓ^* and from $t^*_{\max}$, given that ω^* is selected. For this, we illustrate P_s and D_{95} under the TM 2 for the ALE and PALE configurations in Fig. 4.8; $\ell \in \{8, 16, 32, 64\}$ and $t_{\max} \in \{0.1, 0.2, \ldots, 5\}$ s. Again, only results for $r = 54$ are shown as a similar behavior was observed for $r = 30$. Results obtained with no ACB scheme are also included as a reference.

Fig. 4.8 shows that the P_s obtained with our ACBC scheme is higher than that with no ACB with any $t_{\max} \in \mathbb{R}_{>0}$. It can also be observed that $P_s > 0.95$ for all $t_{\max} > t^*_{\max}$. That is, selecting $t_{\max} > t^*_{\max}$ results in an adequate P_s but slightly increases D_{95}. For example, $D_{95} = 8.380$ and $D_{95} = 8.108$ for the ALE and PALE configurations respectively if an intuitive value $t_{\max} = 1$ s is selected along with $\ell^* = 32$. On the other hand, selecting $t_{\max} < t^*_{\max}$ results in a drastic drop in P_s, except for the ALE configuration with $\ell = 16$. Building on this, and on the fact that in a real world implementation it would be hard to select $t^*_{\max}$ since the exact distribution of the arrivals is ignored, it is advisable to follow a preventive approach and select a relatively high $t_{\max}$.

Also, it can be observed from Fig. 4.8 that the best performance is obtained with $\ell = 32$. That is, the lowest $t^*_{\max}$ was obtained by selecting $\ell = 32$, which leads to the lowest D_{95}. However, the performance obtained with other values of ℓ is just slightly inferior. Consequently, the performance of our ACBC scheme is not greatly affected by the selected value of ℓ, given that excessively short or long values are avoided.

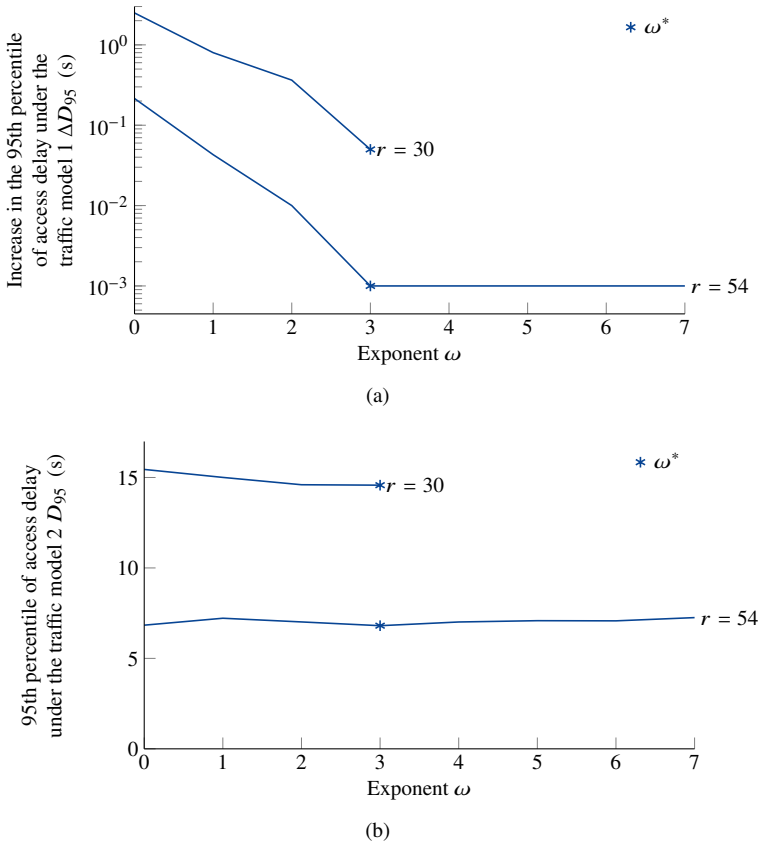


Figure 4.7: (a) Increase in the 95th percentile of access delay under the TM 1 ΔD_{95} and (b) 95th percentile of access delay D_{95} under the TM 2 given $t^*_{\max}$, ℓ^*, and ω for the ALE configuration; $r \in \{30, 54\}$. No $t^*_{\max} \leq 10$ s exists for $\omega \geq 4$ and $r = 30$

4.5.3 Stability test

The analysis of our ACBC scheme concludes with a test of its stability when the mMTC scenario occurs repeatedly over time. For instance, this scenario may arise when a smart metering system, such as the one deployed in a parking lot, is set to shut down when not in use. Therefore, the system will continuously switch between on and off

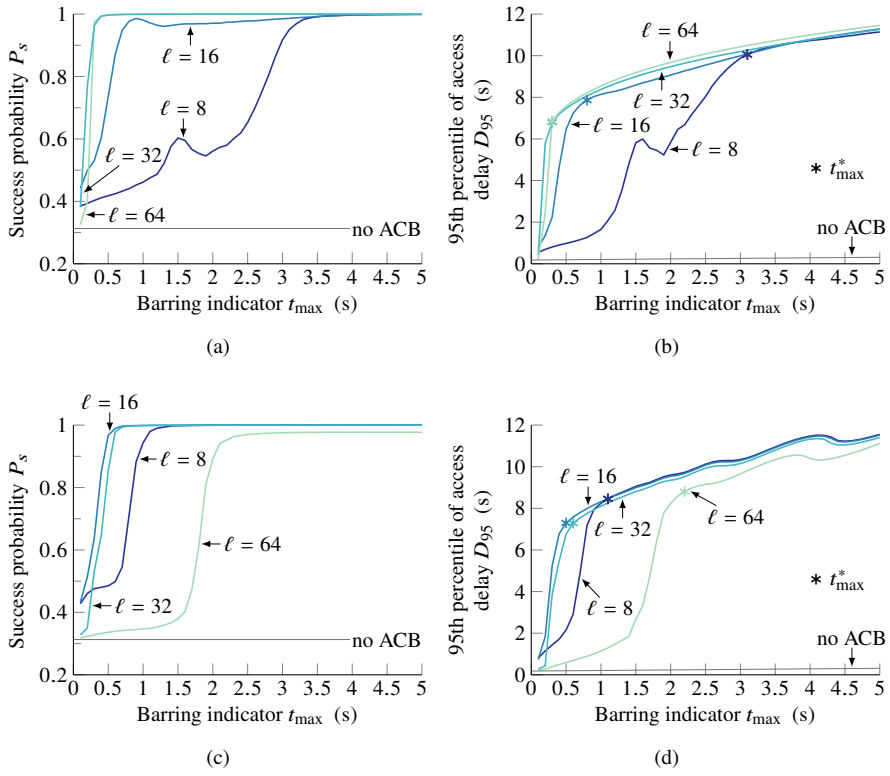


Figure 4.8: Success probability P_s for the: (a) ALE and (b) PALE configurations, and 95th percentile of access delay D_{95} for the: (c) ALE and (d) PALE configurations as a function of t_{max} under the TM 2; $r = 54$ and ω^*.

states; massive accesses will occur every time the system is switched on.

The typical mMTC described by the TM 2 is used for this test. Specifically, the TM 2 is replicated ten times, one after the other. The time between each distribution period is set to 50 s to allow for the stabilization of the filter coefficients; no accesses occur during this period. For this test we follow a different approach before in the sense that we assume an arbitrary $t_{max} = 1$ s is selected in combination with ℓ^* and ω^* (see Table 4.1).

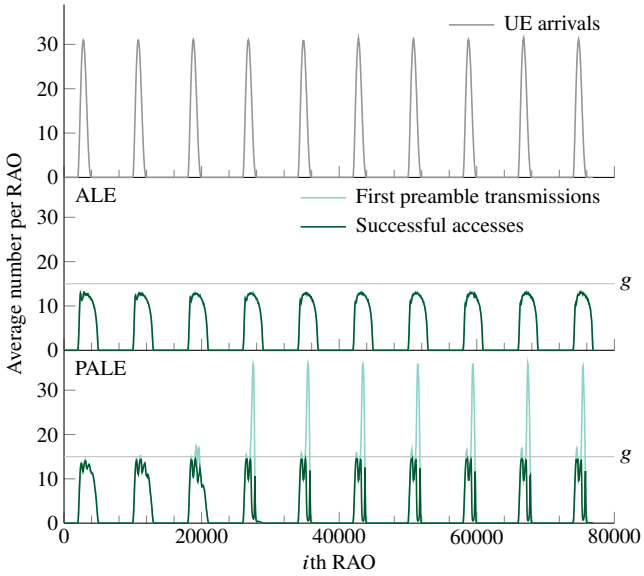


Figure 4.9: Average number of UE arrivals, first preamble transmissions, and successful accesses per RAO for the ALE (middle) and PALE (bottom) configurations under the scenario defined to evaluate the stability of our ACBC.

Fig 4.9 shows the average number of UE arrivals, first preamble transmissions, and successful accesses per RAO for the ALE and PALE configurations under the described scenario and given $r = 54$. As it can be seen, the response of the ALE configuration is adequate for each and every one of the distribution periods. The resulting KPIs with the ALE configuration are: $P_s = 0.997$, $\mathbb{E}[K] = 2.206$, and $D_{95} = 8.302$ s. That is, the success probability is extremely close to one and the D_{95} is less than 22 percent higher than with the optimal configuration. Such a difference should be negligible in the vast majority of delay-tolerant applications.

On the other hand, Fig. 4.9 shows a clear performance degradation with the PALE configuration. Specifically, the average number of successful accesses closely follows that of first preamble transmissions during the first three distribution periods. This is an indicator of an adequate response that is not observed in the rest of the plot. As a result $P_s = 0.646$, which is greatly distant from the target $P_s \geq 0.95$. The main reason

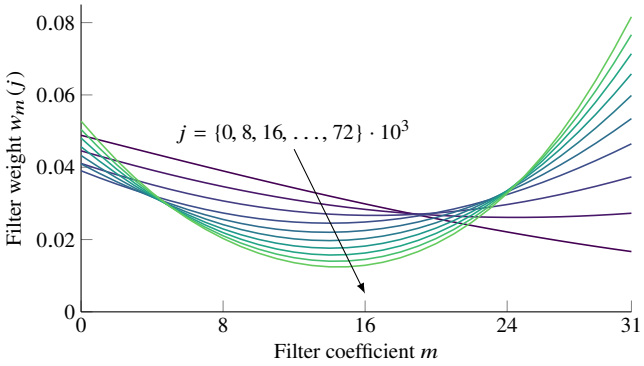


Figure 4.10: Average filter weights $w_m(j)$ for the PALE configuration before the beginning of each distribution period under the scenario defined to evaluate the stability of our ACBC.

for such a great performance degradation is described in the following. The swiftness in the response from the PALE configuration is largely influenced by the filter weights $\boldsymbol{w}$. These wights are adapted by the LMS algorithm to minimize the distance between the desired response and the input to the algorithm. However, there may be several combinations of values of each of the elements in $\boldsymbol{w}$ that minimize this distance under low traffic loads. As a result, the values in $\boldsymbol{w}$ after each distribution period may be different. This is clearly observed in Fig. 4.10, where we show the filter weights before each distribution period.

It is important to observe that the concavity of the plots in this figure increases with time. This phenomenon causes the filter to assign a greater importance to the oldest values of the oldest load indicator (i.e., ratio of idle to available resources) $u(j-\ell+1)$. In other words, the barring rate $p_{\text{acb}}(j)$ is greatly affected by an input that is 2.56 s old. This delay is the result of selecting $\ell = 32$, $t_{\text{si}} = 16$, and $t_{\text{rao}} = 5$ ms. This lack of stability persists even when low signaling traffic loads (in the order of a few accesses per RAO) are injected and is the main drawback of the PALE configuration.

Nevertheless, the identification of this problem in turn allows us to propose some solutions to it. For example, some adaptive algorithms may interpret an abrupt change in the input as a renewed initialization with different initial parameters. In our case,

the high access intensity under the TM 2 represents such a change. The recommended practice in these cases is to reset the filter weights afterwards [45, Chapter 10]. In our case, this would be after each distribution period in which UE arrivals follow the TM 2. This simple solution would make the PALE configuration to "forget" its previous states and to always provide an adequate response from our ACBC scheme.

4.5.4 Impact of realistic assumptions on the performance of the IFI scheme

We conclude our performance analysis by evaluating the performance of the IFI scheme under different scenarios. By doing so, we illustrate the impact of the factors that hinder the accurate selection of barring parameters. We assume the eNB always has perfect information on the number of UEs that will perform an access attempt at each RAO (i.e., even before the RAO occurs), and, hence, can select the optimal barring rate as in (4.15).

The scenarios are defined by two different factors. The first one is the selectivity of the ACB scheme and the second one is the periodicity of the SIB 2 t_{si}. That is, we assume that either: 1) every UE is subject to the ACB scheme; or 2) only the UEs that have not yet begun the RAP are subject to the ACB scheme. We also consider $t_{\mathrm{si}} \in \{1, 16\}$. Hence, we consider the hypothetical case in which $t_{\mathrm{si}} = 1$ besides the lowest achievable $t_{\mathrm{si}} = 16$ RAOs in LTE-A. The combination of these two factors results in the four scenarios included in Table 4.4, where we show the P_s obtained with the IFI scheme.

As it can be seen, the main factor that hinders the proper operation of the IFI scheme is the selectivity of the ACB scheme. That is, the IFI scheme can deal with the real periodicity of the SIB 2 because a sufficiently high $P_s \geq 0.95$ is achieved when every UE is subject to the ACB scheme. On the other hand, the performance of the IFI scheme is poor if the UEs are only subject to the ACB scheme before they initiate the RAP.

Concretely, if we compare the IFI scheme in this latter scenario with the case in which no ACB scheme is implemented (see Table 4.2 on page 105), a similar

Table 4.4: Success probability obtained with the IFI scheme under different scenarios.

UEs subject to the IFI scheme	SIB 2 periodicity t_{si} (RAOs)	Success probability P_s	
		$r = 30$	$r = 54$
Every UE	1	0.987	0.988
	16	0.988	0.971
Only the UEs that have not	1	0.106	0.313
yet begun the RAP	16	0.100	0.312

$P_s = 0.313$ is obtained with $r = 54$. On the other hand, the P_s obtained with $r = 30$ is even lower with the IFI scheme than with no ACB scheme (i.e., $P_s = 0.115$ with no ACB scheme). This problem may be solved by an adequate selection of the barring time $t_{acb}(j)$, but no strategies to fine-tune this parameter were investigated in [35].

4.6 Conclusions

This chapter presented a novel adaptive ACBC scheme; this is our proposed solution support mMTC in 3GPP cellular networks and can be directly implemented at the eNBs. In our ACBC scheme, the selection of the barring parameters is based on the ratio of idle to available resources, which serves as the input to an adaptive filtering process. The LMS adaptive algorithm was selected because of its numerical robustness and simplicity. However, any adaptive algorithm can be selected, but it is also important to recall that initial tests confirmed the superiority of the LMS algorithm when compared to the RLS adaptive algorithm; results with this latter algorithm are included in Appendix C. Therefore, two different configurations for the LMS were analyzed. the first one is the typical ALE and the second one is a novel twist on the latter in which the response is "pulling" toward a desired response. We call this the PALE configuration.

Throughout this chapter we have observed that the target success probability of 95 percent under the TM 2 can be obtained by means of numerous ACBC schemes.

Nevertheless, our ACBC scheme is one of the few that combines the following three characteristics: 1) leads to a nearly optimal performance under periods of no congestion; 2) successfully relieves congestion under mMTC scenarios while obtaining a short access delay; and 3) can be directly implemented in 3GPP cellular networks. And, to the best of our knowledge, is the one that provides the best performance.

Between the ALE and PALE configurations, the latter is the only configuration that minimizes the access delay during intervals of low signaling traffic intensity when $r = 30$. That is, when the RAP capacity is exclusively limited by the number of available preambles. On the other hand, a similar performance can be obtained with both configurations when $r = 54$. That is, when the signaling capacity is limited by the number of available uplink grants. The main practical difference between these configurations relies on the ease of correctly setting the configuration parameters. Concretely, the range of adequate values of $t_{\max}$ is larger for the ALE configuration than for the PALE configuration. On the other hand, the PALE configuration eases the selection of parameters as no exponential ω is needed.

We observed that the stability in the response of the PALE configuration may be compromised if mMTC scenarios occur repeatedly. Nevertheless, this problem can be easily solved by sporadically resetting the filter weights to their original value. On the other hand, results show the stability of the ALE configuration is guaranteed as it is not affected by frequent massive access periods.

Chapter 5

Performance analysis of RA event-reporting in wireless sensor networks (WSNs)

5.1 Introduction

Previous chapters have dealt with massive machine-type communication (mMTC) from the 3rd Generation Partnership Project (3GPP) standard perspective. That is, the performance analysis of the channels involved in the random access (RA) in 4th generation (4G) cellular networks, including the signaling capacity and the access class barring (ACB) scheme, was presented in Chapter 2. Then, an analytical model of the RA in 4G that includes the ACB scheme was presented in Chapter 3. Finally, an adaptive method to fine-tune the ACB parameters was presented in Chapter 4. In this chapter we focus on machine-type communications (MTC) from the wireless sensor network (WSN) perspective. That is, a proprietary solution to relatively small MTC deployments that may be present in urban environments, which aim to improve the quality of life by providing the population with real-time information and services [85].

WSNs are cost-efficient solutions to massive monitoring thanks to its capacity of identifying and reporting a wide range of physical parameters inside the area of interest. A clear example of WSN applications is smart metering, where the network is

in charge of collecting and transmitting environmental parameters such as temperature, pressure, humidity, electrical power, etc.

Advances in the electronics have enabled a sharp increase in the complexity of WSN applications, where multiple physical parameters may be monitored and each of these has its own delay and reliability requirements that must be fulfilled. Specifically, applications that have strict delay and reliability requirements are known as time-critical applications. In these applications, nodes are usually in charge of detecting hazardous conditions, hence, achieving the required report latency is of utmost importance as it allows a proper reaction from the network to the occurring phenomena. In other words, a swift response is needed in order to ensure the timely activation of disaster contention mechanisms and reduce the damage caused to the network, the environment or the population.

The overall behavior of a WSN is determined by the selected WSN protocol, so its selection must be based on the monitoring needs of the WSN user (i.e., owner). WSN protocols are usually a combination of routing and medium access control (MAC) protocols. In the former, nodes are organized depending on their spatial distribution to optimize data transmission paths. *Cluster-based* protocols are a classic form of organization widely used nowadays as they considerably reduce transmission distances and, in turn, energy consumption [13, 115]. In these protocols, nodes are divided in groups named *clusters* during a cluster formation (CF) phase. Each cluster contains a cluster head (CH) node, which is in charge of collecting the data packets from the other nodes in the cluster, cluster members (CMs), and its transmission to the sink node. The latter is a node with higher computational capabilities that gathers the WSN information. It can be, for instance, a personal computer.

MAC protocols, on the other hand, establish the communication links and define the manner in which the nodes share the communication resources. Please observe that the characteristics of the selected MAC protocol must be in line with the requirements of the target application [16, 21, 86]. Otherwise, when the selected protocol is unable to meet the basic application requirements, the whole network is inoperative.

A WSN protocol can perform continuous monitoring, event-driven detection, or both. Continuous monitoring is a proactive approach in which the WSN transmits

environmental information regardless of the state of the physical parameter of interest. On the other hand, event-driven detection is a reactive approach in which the WSN transmits information only when a change in the physical parameter is detected.

My M.Sc. studies were focused on simple WSN protocols with continuous monitoring and event-driven detection capabilities [65]. Then, the beginning of my Ph.D. focused on the analytical modeling of WSN protocols for event-driven detection and reporting in time-critical applications. During these preliminary studies I identified an important aspect that is typically overlooked by MAC protocols for these types of applications is the number of event reports required at the sink node to characterize the occurring phenomena [43]. Naturally, it is not feasible to wait for the reception of the transmissions of every detecting node at the sink. This is simply because WSNs are a distributed entity, and the total number of event detecting nodes is not known prior to the transmission of their packets. Then, a threshold on the number of received event reports must be set to activate the network contention mechanisms in a timely manner. Therefore, the number of required event reports depends on the application. That is, receiving a single event report may be sufficient in certain applications, but in target positioning and tracking applications, which usually use trilateration or triangulation, at least three packets are needed. Besides, when the WSN is in charge of extracting the mobility pattern of targets, the higher the number of transmitted packets, the higher the accuracy [77].

Given the basic application requirements are met, the efficiency of WSN protocols is measured in terms of quality of service (QoS) parameters. Nodes being battery supplied, energy consumption is the QoS parameter most widely studied in the literature [18, 86, 112, 113] as it directly affects network lifetime (period of time for which the network remains functional). The relevance of other QoS parameters such as report latency and event overlooking probability is application dependent. For critical-time applications, these can be equally or even more important than energy consumption.

Despite the high importance of report latency, this parameter is commonly assessed in the literature in terms of its mean value; this is clearly insufficient for time-critical WSN applications. High percentiles or the whole probability distribution of report latency are much better suited in these applications and provide the network adminis-

trator with more meaningful information regarding the behavior of the system than its mean value. Still, the research in this area is scarce.

Building on this, I propose a novel method to calculate the probability distribution of report latency and mean energy consumption in cluster-based RA WSNs. In this method, the probability mass function (pmf) of detecting nodes is first obtained by simulation. Then, discrete-time Markov chains (DTMCs) are used to model the process of RA event reporting. From there, the probability distribution of report latency and the mean energy consumption are obtained. We use this method to evaluate the performance of RA event reporting in applications that require the transmission of a given number of packets to fully characterize the event; a simple RA protocol with overhearing is proposed and evaluated for this kind of applications. That is, neighbors sense the wireless medium to identify when the required number of successful event reports is reached; as it will be observed in Section 5.4 overhearing greatly enhances event reporting when compared to traditional RA, in which every packet is transmitted. Furthermore, we use our method to optimize the QoS of event reporting by identifying the optimal transmission parameters for the selected application prior to network deployment.

One of the main challenges when trying to optimize event reporting WSN protocols is that the optimal transmission parameters depend on the number of detecting nodes for each event occurrence. Naturally, the number of detecting nodes may be different at each event occurrence and is not known before the event detection. In this chapter, we focus on optimizing event report latency by identifying the ideal transmission parameters of detecting nodes in two different approaches: fixed backoff (FB) and adaptive backoff (AB). In the FB, transmission parameters are selected prior to network deployment and remain constant throughout event reporting. In the AB, transmission parameters are selected prior to network deployment and modified by the CMs at each collision.

Results show that increasing the time between successive transmissions, as in the AB can reduce the energy consumption during event reporting when compared to the FB. As an additional benefit, the AB mitigates the negative effects of the inaccurate selection of transmission parameters. This results in a noticeable increase in the

stability of QoS parameters when compared to the FB. These results hold true even when the network operates in a multi-event environment, where it is in charge of monitoring multiple types of events with different characteristics.

The main contributions of this chapter were published in [62], while a preliminary version was published in [63].

5.2 Related work

The importance of the QoS provided by WSN protocols in time-critical applications has increased with recent advances in power electronics [78]. that is, early WSN protocols were specially designed to improve network lifetime [46, 113, 114], and these have served as a base for more complex protocols that aim to reduce report latency and packet loss probability in time-constrained packets. Nevertheless, energy efficiency still is one of the main concerns in WSN, hence these more advanced protocols also have a clear focus on maintaining an adequate energy consumption.

Energy efficiency being so important in WSN, techniques such as sleep scheduling strategies and multi-hop delivery have been proposed to further enhance this parameter [41]. While these techniques reduce energy wastage, they may increase report latency [28, 88, 113] and network congestion [75]. Therefore, sleep scheduling strategies are not optimal for time-critical applications. Other techniques such as dynamic re-routing may be capable of reducing latency and energy consumption but oftentimes rely on a centralized view of the whole network [78], which is not commonly applicable to WSNs.

Hybrid protocols are a distributed energy-efficient solution for time-critical applications [73, 114]. These protocols are capable of adapting its behavior depending on the characteristics of the application and are commonly capable of performing continuous monitoring and event-driven detection simultaneously. The downside of these protocols is, usually, an increase in complexity [64, 65].

In previous studies, we have presented and evaluated the performance of the RA phases of hybrid protocols by means of Markov models [64, 65], but our results were

limited to mean values of energy consumption and report latency. While this is a common practice in the literature [59, 70], it is clearly insufficient for time-critical applications. A feasible option to circumvent this problem in WSN protocols with a scheduled transmission scheme is to assess the report latency by means of its maximum value (i.e., worst-case latency). Such is the case of [117], where a MAC protocol for industrial applications is proposed and analyzed. Nevertheless, assessing report latency in terms of high percentiles or the whole distribution of report latency are much better suited for time-critical applications.

Generic analytic methods for evaluating the performance of RA protocols designed for other types of wireless networks may not be easily adapted for the analysis of cluster-based WSNs due to their particular characteristics. Such is the case of the queuing model for the IEEE 802.11 protocol, presented in [98]. In addition, most of the existing methods for obtaining the probability distribution of report latency in WSNs are protocol specific. The work of Souil *et al.* [92] and of Siddiqui *et al.* [89] are clear examples of these protocol-specific methods. Furthermore, Souil *et al.* and of Siddiqui *et al.* measure report latency as the time required for the first successful transmission to occur; as mentioned in the Introduction, the transmission of a minimum number of packets may be necessary for the accurate characterization of the occurring phenomena. Wang *et al.* [107] are one of the few to consider the need for the transmission of a minimum number of event packets during event reporting and propose a spatio-temporal fluid model, along with a simplified model to obtain the distribution of report delay in multi-hop WSNs. The downside of this model is that its accuracy drops when the node density is low or the traffic rate is high.

At early stages of my studies, we identified the need for a method capable of calculating the report latency distribution for a wide range of WSN protocols and environments. Hence, we developed a hybrid method for obtaining the distribution of report latency. Preliminary results, where only a fixed backoff was considered, were presented in [63]. This latter work was extended to analyze the benefits of adaptive transmission probabilities during RA event reporting WSNs, but also to optimize event reporting. That is, to identify the transmission probabilities that minimize report latency for the fixed and adaptive backoffs, but also reduce energy consumption prior to network deployment.

5.3 Hybrid method for the QoS analysis of RA WSN protocols

In this section we present the network model and the hybrid method for the QoS analysis of RA event reporting WSN protocols. The network model is presented in Subsection 5.3.1, where the network topology, the RA protocol, and the energy consumption model are described. The remaining subsections are dedicated to the describe our hybrid method, which comprises three main phases: 1) obtaining the distribution of detecting CMs; 2) defining the Markov reward process; and 3) obtaining the QoS parameters.

5.3.1 Network model

Following the line of my previous studies [64, 65], we adopt a network topology that has been commonly used in the literature since Heinzelman *et al.* published their paper on the well-known LEACH protocol [46]. In this topology, $m = 100$ nodes are uniformly distributed in a squared area of $100\,\text{m} \times 100\,\text{m}$. That is, from coordinates $(0, 0)$ to $(100, 100)$. The sink node is located outside the supervised area at the coordinates $(200, 0)$.

During event reporting and CF phases, the network operates on a slotted channel, where each time slot is the time required for the transmission of a data or control packet from a CM to the CH and its immediate retransmission to the sink node. The size of the data packet is $l = 2$ kbits, which comprises the data payload, and the identification and type fields. The size of the control packet is $l_c = 1$ kbits, which comprises the same fields but with a shorter payload. The transmission data rate is $R = 40$ kbps and, since two data packets are transmitted one after the other to reach the sink node (i.e., one from the CM to the CH and one from the CH to the sink node) the slot duration during event reporting phases is $t_s = 0.1$ s.

Our study focuses on cluster-based protocols, where the role of nodes acting as either CHs or CMs shifts constantly throughout the operation of the network. This approach avoids the fast battery depletion of nodes acting as CHs. Hence, it reduces

the probability of suffering from the well-known energy hole problem, that may result in some areas being disconnected from the network due to the battery depletion of some critically-placed nodes. Throughout this study we use one of the most important clustering algorithms: LEACH, developed by Heinzelman *et al.* [46]. Through the years, LEACH has served as a base to develop and to assess the efficiency of other routing protocols, this continues to be true even after more than a decade since its publication [13, 28]. However, it is important to emphasize that our hybrid method presents a general structure, so any clustering algorithm can be easily incorporated. As it will be seen in the rest of this chapter, the selected clustering algorithm merely determines the distribution of detecting nodes and clusters; this is obtained by simulation at the first phase of our hybrid model.

The distributed clustering algorithm of the LEACH protocol is performed as follows. The operation of the network is divided into rounds and a CF phase is performed at the beginning of each round. At each CF phase, each node has a certain probability of being elected as a CH; this probability increases with the number of rounds. Once a node has been elected as a CH, the probability of being elected as a CH once more in the next few rounds becomes 0. Afterwards, the CHs inform their status to the other nodes by broadcasting an advertisement message (using CDMA). The remaining nodes join a given cluster based on the received signal strength of the CH transmission; this is, typically, the one with the closest CH. The interested reader is referred to [46] for more details on the LEACH protocol.

The energy to receive a packet depends on its length and on the energy required per bit by the communication circuits E_{elec} as

$$E_{\text{rx}}(l) = lE_{\text{elec}}. \tag{5.1}$$

In this study we adopt a generic energy consumption model that is widely used in the literature, in which $E_{\text{elec}} = 50\,\text{nJ/bit}$ [28, 46, 114]; though any energy consumption model can be incorporated to our hybrid method.

During random access, the communication circuits in every CH must be active whenever CM transmissions can occur, so they are able to relay any received data packet containing an alarm or control message from its CMs to the sink node with

minimal latency. The energy required to transmit a packet depends on both the length of the packet l and the selected transmission range d m [46]. Therefore, the total energy consumed during a packet transmission is given as

$$E_{\text{tx}}(l, d) = lE_{\text{elec}} + l\epsilon_{\text{amp}}d^{p_l}, \tag{5.2}$$

where the energy required per bit and per square meter by the transmission amplifier is $\epsilon_{\text{amp}} = 10\,\text{pJ/bit/m}^2$ and p_l is the path loss exponent.

Two power levels are defined for packet transmissions: low and high power. Low power transmissions consume E_{cmtx} J and are used for CM to CH communication. By using this power level, nodes are able to perform transmissions at up to $d_l = 35$ m. High power transmissions consume E_{chtx} J and are used for CH to sink communication. By choosing this power level, nodes are able to transmit from the farthest possible coordinates within the network to the sink node, which is $d_h = \sqrt{200^2 + 100^2}$ m. This approach eliminates the need of calculating the minimum energy required for transmission; this latter approach is used by Heinzelman *et al.* [46]. An additional benefit of defining these two power levels is that packet loss probability due to changes in the wireless environment is considerably reduced when compared to the minimum energy approach. For instance, the minimum energy required for a successful packet transmission may vary through time due to changes in the wireless channel conditions, so it must be calculated periodically. Furthermore, event-driven detection WSN protocols are characterized by infrequent transmissions. For instance, the frequency of occurrence of relevant events may range from once a day for highly frequent events, to once a month for relatively infrequent events, or even to once a year for rare events. Therefore, energy efficiency is not significantly affected by following this approach when compared to the minimum energy approach.

Events are generated as in Calafate *et al.* [25], where a model for indoor gas propagation is presented. Each event has originating coordinates (w, h), which are selected by means of a two-dimensional uniform random variable (RV). That is, $w, h \equiv \text{U}[0, 100]$; w and h are selected independently. A node detects an event when the reading of the phenomena of interest causes the reading in one or more of its sensors to exceed a threshold. In our case, we assume the sensitivity of the sensors is such

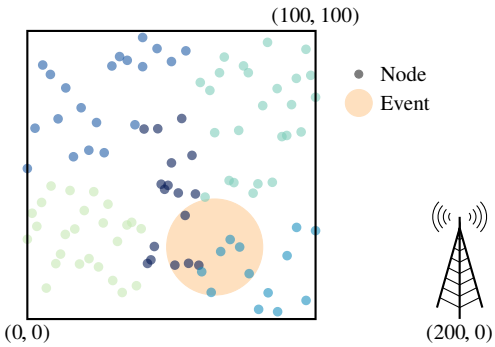


Figure 5.1: Example of a cluster-based WSN for event detection and reporting. Nodes detecting the event will transmit a data packet to their CH. Different colors indicate different clusters.

that the threshold is exceeded when the node is located within $r \in \{5, 10, 15, 20, 30\}$ m from the event originating coordinates.

In order to evaluate the performance of RA protocols in extreme conditions, it is assumed that the event is detected simultaneously by all the CMs within a radius r from the event originating coordinates. On the other hand, the event is not detected by CMs located at a radius greater than r from the event originating coordinates. We assume that only one packet is generated per node per detected event. This behavior can be achieved, for example, by implementing a double sliding window scheme [42]. Furthermore, we do not consider the possibility that an event occurs while the reporting of a previous event is still ongoing in the same cluster. Please observe that this is a valid approach because, due to the nature of the events, this scenario has an extremely low probability of occurrence.

Fig. 5.1 summarizes the considered scenario. In this figure, nodes are uniformly distributed in the area of interest. As indicated by the different colors in Fig. 5.1, nodes form five clusters. Then, an event occurs and is detected by several nodes; in this case, the event is detected by nodes from two different clusters. Detecting nodes will transmit a data packet to their CH, which is in charge of relaying this packet to the sink node.

Given the random nature of the event occurring coordinates and of the network topology, the number and the spatial distribution of detecting nodes may be different for each event occurrence. Building on this, we define RV N_{tot} as the total number of detecting CMs, RV N_c as the number of clusters with detecting CMs, and RV N_i as the number of detecting CMs in the ith cluster. That is, each detecting cluster is assigned an index $i \in \{1, 2, \ldots, N_c\}$ that serves as a temporal identifier. In addition, to simplify notation we define RV N as the number of detecting CMs at a given cluster, whose support is $n \in \{\mathbb{Z}_+ \mid n < m\}$, the number of nodes.

In the following, the RA protocol used for event reporting is described. Please observe a similar RA protocol is implemented for CF phases. Upon the occurrence of a new event, detecting nodes attempt transmission with probability τ per time slot and backoff is used for collision handling. That is, the transmission probability of the CMs for the first transmission attempt is τ at each time slot. Then, the transmission probability for collided CMs (CBMs) becomes $\beta = \tau/b$ for the subsequent transmissions, where $b \geq 1$. Clearly, $b = 1$ and $b > 1$ correspond to the FB and AB approaches, respectively. Whenever the CH receives an event packet, it is directly transmitted to the sink node in the same time slot (i.e., second half) in order to minimize report latency. At the end of event reporting, nodes reset their transmission probabilities to the initial value τ.

Code Division Multiple Access (CDMA) is used to avoid inter-cluster collisions. For this, a CDMA code is selected per cluster and used for transmissions from CMs to the CH and from the CH to sink [48]. Consequently, it is safe to assume that, during CF and event reporting phases, collisions can only occur between CMs from the same cluster. In other words, we consider each cluster operates independently during RA. As such, we analyze event reporting in terms of independent clusters by means of a DTMC.

Next, let k be the number of messages required to fully characterize the occurring phenomena. Once the sink node receives k event messages, it reacts accordingly. Conversely, in cases where $k > N_{\text{tot}}$, the event is overlooked. During event reporting, the network can be configured to transmit either N or k data packets per cluster. In the former, every detecting CM is set to transmit its data packet, which results

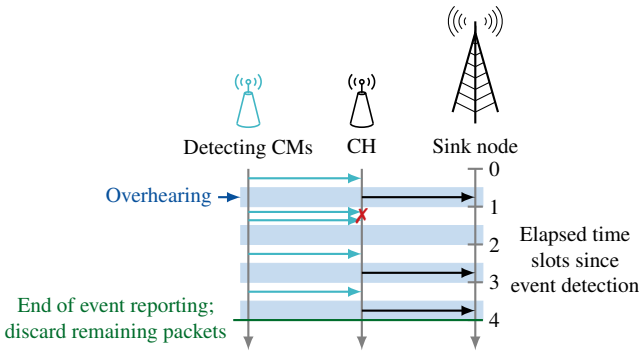


Figure 5.2: RA event reporting in time-critical applications with overhearing for $k = 3$.

in N_i transmissions to the sink node from the ith cluster. In the latter, CMs sense (i.e., overhear) the wireless medium during event reporting until the kth packet is successfully transmitted. At this point in time, the remaining $N_i - k$ CMs in the ith cluster discard their data packets, which reduces energy wastage. Fig. 5.2 illustrates the process of event reporting for the considered set of applications.

It is worth noting that a consequence of using CDMA to avoid inter-cluster collisions is that CMs are only aware of successful transmissions within their own cluster. As a result, event reports can be restricted within clusters but not within the entire network. Consequently, whenever each cluster is set to send k packets and an event is detected in $N_c > 1$ clusters, more than k packets can be received at the sink node.

5.3.2 Obtaining the distribution of detecting CMs

The probability distribution of the number of nodes that detect the event simultaneously is the input to our analytical model. Hence, its adequate calculation is mandatory to effectively calculate the QoS parameters and to optimize the RA protocol. Since we focus on cluster-based WSN protocols and each cluster operates independently during event reporting (due to the use of CDMA), we obtain the distribution of the number of detecting clusters (i.e., clusters with at least one detecting node) N_c and of detecting CMs per cluster N_i for $i \in \{1, 2, \ldots, N_c\}$ by simulation. For this, we have

developed a discrete event simulator in C, in which the selected clustering algorithm is implemented, along with the generation and detection of events. As mentioned at the beginning of this section, the LEACH protocol is used throughout this chapter, but any clustering algorithm can be selected.

In each simulation, nodes are first randomly distributed within the area of interest; their coordinates are selected by generating two random numbers with uniform distribution independently for each node, namely h and w. Then, nodes are organized in clusters according to the selected clustering protocol. Next, 1000 events are generated after each CF phase and the number of detecting nodes is obtained and stored. A total of 20 CF phases are performed for each simulation run, hence, 20 000 events are generated per simulation.The number of simulation runs is such that the error between the cumulative results (i.e., the distribution of detecting CMs) obtained until the j simulation differ from those obtained at the $(j-1)$th simulation by less than 10^{-5}.

Fig. 5.3 and Fig. 5.4 present the results obtained by simulation. Specifically, Fig. 5.3 shows the cumulative distribution function (CDF) of the number of detecting nodes for a given number of detecting clusters. That is, $F_{N|N_c}(n)$ for $N_c \in \{1, 2, 3, 4\}$ given $r \in \{5, 10, 15, 20, 25, 30\}$ m. Naturally, the greater the number of detecting clusters, the smaller the number of detecting CMs per cluster.

Next, Fig 5.4 shows the CDF of the number of detecting clusters and, as expected, this number grows with the event detection radius. Still, it is important to observe that, for the selected topology, more than 97 percent of the events with $r = 5$ are detected in only one cluster. On the other hand, events are rarely detected in 5 or more clusters, even when the largest event radius $r = 30$ is selected.

The behavior depicted in Fig. 5.3 and Fig. 5.4 has especial significance when calculating the event overlooking probability $\Pr[N_{\text{tot}} < k]$. That is, the probability that less than k nodes detect an occurring event. In such cases event reporting is unsuccessful as the sink is unable to obtain the necessary information to characterize the event.

Event overlooking probability can be easily calculated with results obtained from simulation and is shown in Fig. 5.6. Naturally, this probability is sharply reduced as r increases. Nevertheless, as the number of detecting nodes sharply increases with

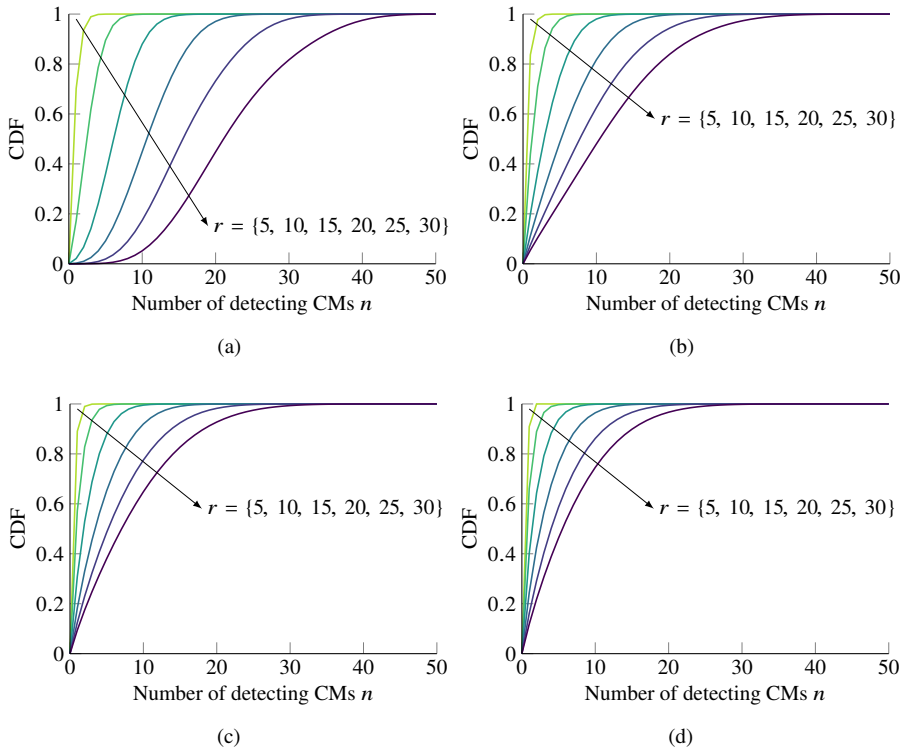


Figure 5.3: CDF of the number of detecting CMs per cluster, N, given the event is detected in N_c clusters for $r \in \{5, 10, 15, 20, 25, 30\}$ m and: (a) $N_c = 1$, (b) $N_c = 2$, (c) $N_c = 3$, and (d) $N_c = 4$.

r, the chances of network congestion in RA protocols also increase. Therefore, the network administrator must configure the thresholds in the nodes' sensors to achieve a sufficiently long detection radius, and hence, a sufficiently low event overlooking probability. But also, r should be sufficiently short to avoid excessive event detections.

Clearly, the lower limit for the event detection radius highly depends on the density of deployed nodes within the network and on the requirements of the application. For the considered node density of 0.01 nodes/m^2, and $k = 3$, we assume that $r \geq 15$ m

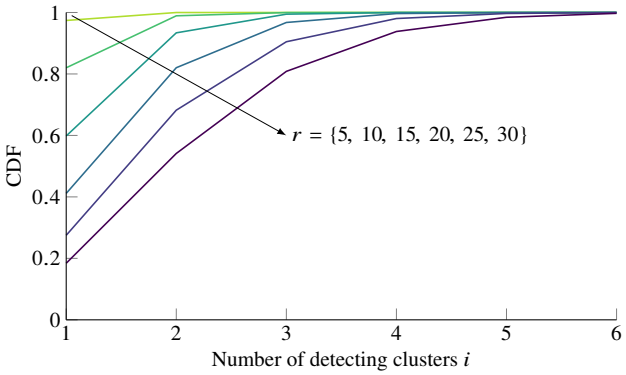


Figure 5.4: Pmf of the number of detecting clusters N_c for $r \in \{5, 10, 15, 20, 25, 30\}$ m.

results in an adequate $\Pr[N_{\text{tot}} < 3] < 0.09$. Hence, $k = 3$ will be selected throughout Section 5.4 and the focus will on $r \geq 15$ m. Please observe the selected value of $k = 3$ goes in line with the requirements of target positioning and tracking applications. That is, these usually use trilateration or triangulation, so at least three packets are needed. Building on this, in Section 5.4 we focus on assessing and optimizing the energy consumption and report latency for large event detection radii. That is, to identify the transmission probabilities that optimize performance given a sufficiently large r has been selected.

5.3.3 Defining the Markov reward process

As mentioned above, the use of CDMA allows us to analyze the system in terms of independent clusters. For this, we seek to define a DTMC that describes the process of event reporting within each cluster. As a starting point, we present the Markov model for the FB approach. The resulting DTMC is depicted in Fig. 5.6. This model shares some similarities with those used in our previous studies [63–65], where the transmission probability remains unaffected during backoff (i.e., $b = 1$). But also presents an important difference that is described in the following.

The model starts at state $(N = n)$, where n is the number of CMs that have detected

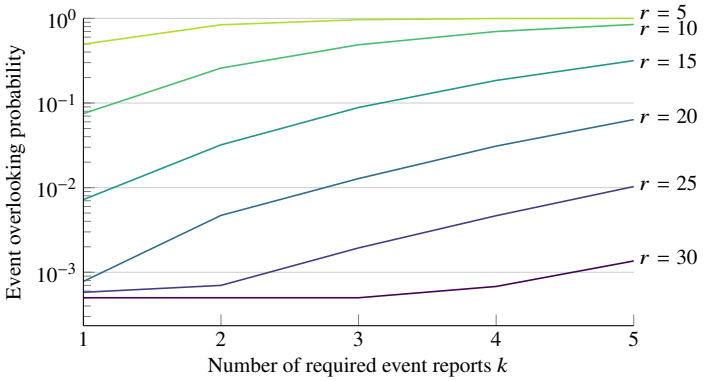


Figure 5.5: Event overlooking probability for the considered node density of 0.01 nodes/m^2, $k \in \{1, 2, \ldots, 5\}$, and $r \in \{5, 10, 15, 20, 25\}$.

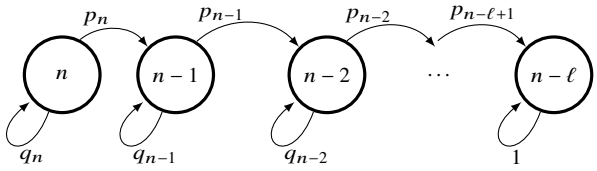


Figure 5.6: DTMC that describes the random access event reporting over a slotted channel with the FB.

the event in a given cluster. That is, n is the outcome of a single experiment for RV N. As such, n can be different for each event. Therefore, the state-space of the DTMC is $\mathscr{S} = \{x \in \mathbb{N} \mid n - \ell \leq x \leq n\}$; x represents the number of CMs with pending event transmissions. At each time slot, the DTMC can either transition towards the absorbing state $(n - \ell)$, where $\ell = \min\{k, n\}$, or remain in the same state. Please mind the use of variable ℓ instead of k. The reason for this is that the number of detections within a cluster n can be less than k. Still, these cases contribute to event reporting within the network; this is a critical difference with respect to models used in previous studies.

A transition from an arbitrary transient state (x) to $(x - 1)$ occurs with probability

p_x; this is the probability of a successful transmission, which occurs whenever a single CM sends a data packet. Conversely, the probability of remaining in the same state is the probability of an unsuccessful event report $q_x = 1 - p_x$. This occurs either when none of the CMs attempts transmission or when a collision occurs.

Next, let $S(x)$ be the RV that defines the number of event transmissions within a cluster at an arbitrary time slot for a given x. Building on this, a successful transmission occurs with probability

$$p_x = \Pr[S(x) = 1] = x\tau(1-\tau)^{x-1}. \tag{5.3}$$

The absorbing DTMC that describes the process of event reporting (i.e., the transmission of the first $\ell = \min\{k, N\}$ messages within a cluster) is depicted in Fig. 5.6. Then, the substochastic matrix that represents the transitions within transient states is

$$\boldsymbol{T} = \begin{bmatrix} q_n & p_n & 0 & \cdots & 0 \\ 0 & q_{n-1} & p_{n-1} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & p_{n-\ell+2} \\ 0 & 0 & 0 & \cdots & q_{n-\ell+1} \end{bmatrix}$$

Also, from Fig. 5.6 we derive a Markov reward process that allows us to calculate the mean energy consumption during event reporting $\overline{E}$. For this, rewards are given as the mean energy consumption at each system transition. In basic RA (no medium sensing) successfully transmitting an event packet requires a CM to CH (E_{cmtx}) and a CH to sink (E_{chtx}) transmission. Thus, the transition from any arbitrary transient state (x) to ($x - 1$) has a reward

$$\rho'(p_x) = E_{\text{cmtx}} + E_{\text{chtx}}. \tag{5.4}$$

In case no transmission is attempted by the CMs or a collision occurs, the system remains in the same state. As a result, the reward for remaining in state x is given as the energy consumed by a CM transmission times the expected number of transmissions, given none or multiple transmissions occurred

$$\rho'(q_x) = E_{\text{cmtx}}\mathbb{E}[S(x) \mid S(x) \neq 1] = \frac{x\tau - p_x}{1 - p_x} E_{\text{cmtx}}. \tag{5.5}$$

In order to identify the time slots in which successful and failed transmissions occur within the cluster, the CMs must be set to overhear CH transmissions during event reporting. This enables the CMs to discard any remaining event packets once k messages have been successfully transmitted. Therefore, the reward for the transition from state (x) to $(x-1)$ becomes

$$\rho(p_x) = \rho'(p_x) + (x-1)E_{\text{elec}} = E_{\text{cmtx}} + E_{\text{chtx}} + (x-1)E_{\text{elec}} \tag{5.6}$$

and remaining in the same arbitrary transient state x has a reward given as

$$\rho(q_x) = \rho'(q_x) + (x - \mathbb{E}[S(x) \mid S(x) \neq 1])E_{\text{elec}} = \frac{(E_{\text{cmtx}} - E_{\text{elec}})(x\tau - p_x)}{1-p_x} + xE_{\text{elec}}. \tag{5.7}$$

We use these rewards to construct the reward matrix

$$\boldsymbol{R} = \begin{bmatrix} \rho(q_n) & \rho(p_n) & 0 & \cdots & 0 \\ 0 & \rho(q_{n-1}) & \rho(p_{n-1}) & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \rho(p_{n-\ell+2}) \\ 0 & 0 & 0 & \cdots & \rho(q_{n-\ell+1}) \end{bmatrix}$$

This concludes the model for the FB approach. In the following, we present the Markov model for the AB approach.

Given the AB is implemented, N nodes initiate the event reporting process with transmission probability τ. Then, whenever a collision occurs, the transmission probability of implicated CMs becomes $\beta = \tau/b$, where $b \geq 1$. The rationale behind this approach is simple: collisions indicate that several CMs are indeed competing for medium access. Hence, reducing the transmission probabilities decreases collision probability; this in turn increases the probability of successful transmissions. Please observe that following the opposite approach (i.e., $b < 1$) will rarely enhance the performance of RA protocols, while $b = 1$ corresponds to the FB approach. The number of CMs that have caused a collision and perform backoff, hereafter denoted as backoff CMs (BCMs), is z. Once event reporting is concluded, the transmission probability of CMs is reset to its original value τ. As a summary, the Markov model that is presented in the following is a generalization of the model for the AB, where $b = 1$.

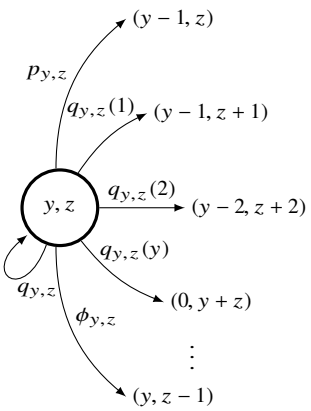


Figure 5.7: Possible transitions of the two-dimensional DTMC from an arbitrary transient state (y, z).

In this model, the states of the DTMC are determined by the number of detecting CMs that have not yet attempted transmission y and by the number of BCMs z. In other words, the state-space of the DTMC for the AB is $\mathscr{S} = \{y, z \in \mathbb{N} \mid n - \ell \leq y + z \leq n\}$; hence the number of CMs with pending transmissions is $x = y + z$. The transmission probabilities of the y and z CMs are τ and β, respectively. The DTMC begins at state $(N = n, 0)$. that is, N CMs detecting the event and 0 BCMs with reduced transmission probability, and evolves towards state $(0, 0)$.

Fig. 5.7 shows every possible transition from an arbitrary transient state (y, z). Please observe now transitions depend on the number of CMs and BCMs that perform a transmission at a given time slot, defined by $S(y)$ and $S(z)$, respectively, but also on their transmission probabilities τ and β, respectively.

Specifically, transitions from the arbitrary transient state (y, z) to $(y - 1, z)$ occur with probability

$$p_{y,z} = \Pr\left[S(y) = 1\right] \Pr\left[S(z) = 0\right] = y\tau(1 - \tau)^{y-1}(1 - \beta)^{z}, \tag{5.8}$$

which represents a successful transmission from one out of the y CMs. Transitions

from state (y, z) to $(y, z-1)$ occur with probability

$$\phi_{y,z} = \Pr\left[S(y) = 0\right] \Pr\left[S(z) = 1\right] = (1-\tau)^y z\beta(1-\beta)^{z-1}, \tag{5.9}$$

which represents a successful transmission from one out of the z BCMs. Therefore, the probability of a successful transmission is

$$p_{y,z} + \phi_{y,z}. \tag{5.10}$$

Conversely, the probability of remaining in the same state is

$$q_{y,z} = \Pr\left[S(y) = 0\right] \Pr\left[S(z) \neq 1\right] = (1-\tau)^y \left(1 - z\beta\,(1-\beta)^{z-1}\right), \tag{5.11}$$

which occurs whenever none of the y CMs and none or multiple of the z BCMs transmit.

Finally, transitions from (y, z) to $(y-v, z+v)$ denoted as $q_{y,z}(v)$, where v represents the number of CMs that become BCMs, are divided in two cases: $v = 1$ and $v \geq 2$. In the former, one out of the y CMs and at least one of the z BCMs attempt transmission, which occurs with probability

$$q_{y,z}(1) = \Pr\left[S(y) = 1\right] \Pr\left[S(z) \neq 0\right] = y\tau(1-\tau)^{y-1}\left(1 - (1-\beta)^z\right) \tag{5.12}$$

In the latter, $v \geq 2$ of the y CMs are involved in a collision, which occurs with probability

$$q_{y,z}(v) = \Pr\left[S(y) = v \mid v \geq 2\right] = \binom{y}{v} \beta^v\,(1-\beta)^{y-v}, \qquad v \geq 2. \tag{5.13}$$

Please observe that the number of transmitting BCMs $S(z)$ is irrelevant for $q_{y,z}(v)$ when $v \geq 2$.

At this point we have defined the probability of every possible transition. With this information, we build the substochastic matrix that represents the transitions between transient states as follows.

$$\boldsymbol{T}_a = \begin{bmatrix} q_{n,0} & 0 & q_{n,0}(2) & \cdots & 0 \\ 0 & q_{n-1,1} & q_{n-1,1}(1) & \cdots & 0 \\ 0 & 0 & q_{n-2,2} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & q_{1,n-\ell}(1) \\ 0 & 0 & 0 & \ldots & q_{0,n-\ell+1} \end{bmatrix}$$

The dimension of this square substochastic matrix depends on the initial number of detecting nodes for a specific experiment n and k. Specifically, the number of rows/columns of matrix $\boldsymbol{T}_a$ is

$$\sum_{j=0}^{\ell-1} n+1-j = \ell\,(n+1) - \sum_{j=0}^{\ell-1} j. \tag{5.14}$$

That is, state $(n, 0)$ has $n+1$ possible transitions to states in which the total number of contending CMs remains unaffected and only transition to state $(n-1, 0)$ reduces the total number of contending CMs by 1; this transition occurs with probability $p_{n,0}$.

From there, we build the reward matrix $\boldsymbol{R}_a$ in the same manner as $\boldsymbol{R}$, where the reward for a successful transmission at an arbitrary state (z, y) is

$$\rho_a(p_{y,z}) = \rho_a(\phi_{y,z}) = E_{\text{cmtx}} + E_{\text{chtx}} + (y+z-1)E_{\text{elec}}. \tag{5.15}$$

The reward for remaining at the same state, that is, when none of the y CMs and either none or more than two of the z BCMs perform a transmission is given as

$$\rho_a(q_{y,z}) = \frac{(E_{\text{cmtx}} - E_{\text{elec}})(z\beta - \phi_{y,z})}{1 - \phi_{y,z} + xE_{\text{elec}}} \tag{5.16}$$

and the reward for transition from the state (y, z) to $(y-v, z+v)$ is

$$\rho_a(q_{y,z}(v)) = \rho_a(q_{y,z}) + v\,(E_{\text{cmtx}} - E_{\text{elec}})\,. \tag{5.17}$$

Building on this, the resulting reward matrix is

$$\boldsymbol{R}_a = \begin{bmatrix} \rho_a\,(q_{n,0}) & 0 & \rho_a\,(q_{n,0}(2)) & \cdots & 0 \\ 0 & \rho_a\,(q_{n-1,1}) & \rho_a\,(q_{n-1,1}(1)) & \cdots & 0 \\ 0 & 0 & \rho_a\,(q_{n-2,2}) & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & \rho_a\,(q_{1,n-\ell}(1)) \\ 0 & 0 & 0 & \ldots & \rho_a\,(q_{0,n-\ell+1}) \end{bmatrix}$$

As in the transition matrix, we generate a different reward matrix for each possible value of N. By building these matrices, we are now able conduct the performance analysis of event reporting for any given value of $b \in [1, \infty)$.

5.3.4 Obtaining the QoS parameters

Once we have constructed the transition $\boldsymbol{T}$ and reward $\boldsymbol{R}$ matrices, we proceed to calculate the energy consumption during event reporting. As a starting point, we obtain the energy consumption in each arbitrary transient state for the RA protocol with FB (i.e., $b = 1$) by solving the following set of Bellman equations [23].

$$\overline{E}_x = q_x \left[r(q_x) + \overline{E}_x \right] + p_x \left[r(p_x) + \overline{E}_{x-1} \right]. \tag{5.18}$$

That is, (5.18) can be solved either as a set of linear equations or recursively for all possible values of the initial number of detecting CMs n, given the initial condition $\overline{E}_{n-\ell} = 0$. That is, the energy consumption at the absorbing state is 0 J. For the AB where $b \neq 1$, the energy consumption in each arbitrary transient state (y, z) is calculated similarly

$$\begin{aligned} \overline{E}_{y,z} =& q_{y,z} \left[\rho(q_{y,z}) + \overline{E}_{y,z} \right] + p_{y,z} \left[\rho(p_{y,z}) + \overline{E}_{y-1,z} \right] \\ &+ \phi_{y,z} \left[\rho(\phi_{y,z}) + \overline{E}_{y,z-1} \right] + \sum_{j=1}^{y} q\,(j)_{y,z} \left[\rho(q_{y,z}\,(j)) + \overline{E}_{y-j,z+j} \right]. \end{aligned} \tag{5.19}$$

Please recall that our first step was to obtain the pmf of the number of clusters with detecting CMs $\Pr[N_c = i]$ and of the number of detecting CMs given N_c detecting clusters $\Pr[N = n \mid N_c = i]$ by simulation. With this information, the mean energy consumption during event reporting in a given scenario, denoted as $\overline{E}$, can be obtained as

$$\overline{E} = \sum_{i=1}^{i_{\max}} i \Pr[N_c = i] \sum_{n=1}^{n_{\max}} \Pr[N = n \mid N_c = i]\, \overline{E}_n, \tag{5.20}$$

where $n_{\max}$ and $i_{\max}$ are the maximum values of n and i obtained by simulation. From there, the mean energy consumption for the case where $b \neq 1$ is easily calculated by substituting $\overline{E}_n$ with $\overline{E}_{n,0}$ in (5.20), since $z = 0$ at the beginning of event reporting.

Next, we calculate the event report latency. For this, let T_ℓ be the RV that defines the number of time slots elapsed between the occurrence of an event, detected by N CMs, and the end of the event reporting process in a single cluster. Please recall that event reporting in a given cluster is completed when $\ell = \min\{k, n\}$ out of a total $N = n$ packets are transmitted successfully from the detecting CMs to the sink node. As

such, T_ℓ is the time to absorption in one of the DTMCs defined earlier in this section; absorption occurs when the ℓth packet is successfully transmitted.

Therefore, T_ℓ has a phase-type (PH) distribution with representation $(\boldsymbol{\alpha}, \boldsymbol{T})$, where $\boldsymbol{\alpha}$ is the row vector of initial probabilities and $\boldsymbol{T}$ is the substochastic matrix of transient states. In other words, $\boldsymbol{\alpha}$ defines the probabilities that the system starts at each of the transient states. Also let $\mathbf{1}$ be a column vector of 1s of the same dimension as $\boldsymbol{T}$.

For the FB approach depicted in Fig. 5.6, $\boldsymbol{\alpha} = [1\ 0\ \ldots\ 0]$ of length ℓ. On the other hand, for the AB approach depicted in Fig. 5.7, the length of vector $\boldsymbol{\alpha}$ is the number of rows in matrix $\boldsymbol{T}_a$, defined in (5.14). Once $\boldsymbol{\alpha}$ is defined, the expected value of T_ℓ can be easily obtained as

$$\overline{T}_\ell = \boldsymbol{\alpha}(\boldsymbol{I} - \boldsymbol{T})^{-1}\mathbf{1}, \tag{5.21}$$

where $\boldsymbol{I}$ is the identity matrix of the same dimension as $\boldsymbol{T}$ [17]. Then, let T be the RV that defines the report latency for a given scenario, described by $\Pr[N_c = i]$ and $\Pr[N = n \mid N_c = i]$. Its mean value can be calculated as

$$\overline{T} = \sum_{i=1}^{i_{\max}} \Pr[N_c = i] \sum_{n=1}^{n_{\max}} \Pr[N = n \mid N_c = i]\, \overline{T}_\ell. \tag{5.22}$$

However, as stated above, we are especially interested in obtaining the probability distribution of the report latency T. For this, Let s be the number of elapsed time slots since the detection of the event. Next, let $\{X_s(c)\}_{s\in\mathbb{Z}_+}$ be the stochastic process (i.e., collection of RVs) that defines the total number of packets that are successfully transmitted from c clusters to the sink node s time slots after the detection of the event. Therefore, the support of RV $X_s(c)$ is $j \in \{0, 1, \ldots, n_{\max}\}$.

Then, the pmf of RV $X_s(c)$ conditioned to the number of detecting clusters i can be determined by means of the following recursion.

$$\begin{aligned} p_{X_s}(j, c \mid i) &= \Pr\left[X_s(c) = j \mid N_c = i\right] \\ &= \sum_{u=0}^{j} p_{X_s}(u, c-1 \mid i)\, p_{X_s}(j-u, 1 \mid i), \\ &\qquad \text{for } c \in \{1, 2, \ldots, i\} \text{ and } j \in \{0, 1, \ldots, n_{\max}\} \end{aligned} \tag{5.23}$$

where

$$p_{X_s}(j, 1 \mid i) = \sum_{n=1}^{n_{\max}} \Pr[N = n \mid N_c = i] \Pr[X_s(1) = j] \tag{5.24}$$

is the probability that j packets are successfully transmitted within one cluster in s time slots or less after an event is detected in i clusters.

But we are still missing $\Pr[X_s(1) = j]$; the remaining piece of information to calculate the pmf of T. However, it can be easily obtained from the vector of initial states $\boldsymbol{\alpha}$ and matrix $\boldsymbol{T}$ as follows.

$$\boldsymbol{\alpha}^{(s)} = \boldsymbol{\alpha}^{(s-1)}\boldsymbol{T} \tag{5.25}$$

given the initial condition $\boldsymbol{\alpha}^{(0)} = \boldsymbol{\alpha}$.

Then, for the FB approach depicted in Fig. 5.6 we have

$$\Pr[X_s(1) = j] = \alpha_j^{(s)} = \boldsymbol{\alpha}^{(s)}\boldsymbol{T}\boldsymbol{e}_j, \qquad \text{for } j \in \{1, 2, \ldots, \ell - 1\} \tag{5.26}$$

where $\boldsymbol{e}_j$ is a column vector of length ℓ; this vector has a 1 at position $j + 1$ and the rest of the entries are 0s. This probability is obtained analogously for the AB approach depicted in Fig. 5.7; the only difference is the number of entries of the vectors involved. Please observe that (5.25) and (5.26) are different to the equations presented in our previous work [62]. The reason for this is that the formulas presented in this chapter are much more computationally efficient than those used in prior work.

We proceed to obtain the probability that the system is at any of the transient states at time index s,

$$\Pr[T_\ell > s] = \boldsymbol{\alpha}^{(s)}\boldsymbol{T}\mathbf{1}. \tag{5.27}$$

Naturally, $\Pr[T_\ell > s] = 1$ if $s \leq \ell - 1$. Building on this, we obtain the CDF of T_ℓ as follows.

$$F_{T_\ell}(s) = 1 - \Pr[T_\ell > s] = \Pr[X_s(1) = \ell]. \tag{5.28}$$

Finally, when the number of event packets required at the sink is k, the CDF of T given the number of affected clusters is $N_c = i$ is

$$\Pr[T \leq s \mid N_c = i; k] = \sum_{j \geq k} p_{X_s}(j, i \mid i) = 1 - \sum_{j < k} p_{X_s}(j, i \mid i). \tag{5.29}$$

Finally, we can obtain the CDF of report latency for a specific environment as

$$F_T(s) = \Pr[T \leq s; k] = \sum_{i=1}^{i_{\max}} \Pr[T \leq s; N_c = i, k] \Pr[N_c = i], \tag{5.30}$$

given the distribution of N and N_c are known. In the following section, we study the impact that transmission probabilities and event detection radii have on performance.

5.4 QoS analysis

The present section has been divided in three main subsections. The first two sections are dedicated to the FB and AB approaches, respectively, under environments in which a single type of event occurs. That is, every occurring event has the same detection radius r. The third subsection presents results derived from environments where two type of events occur; these two types of events have different detection radii r.Results presented in this section highlight the capabilities of the proposed method for QoS analysis and also showcase the robustness of the AB to the inadequate selection of parameters.

The relevant network parameters used throughout this section are listed in Table 5.1.

Please recall that at the beginning of the previous section we performed a study to determine the event overlooking probability; results were presented in Fig. 5.5. That is, in cases where $n_{\text{tot}} < k$, the sink is unable to obtain the necessary information for characterizing the event. Hence, event reporting is unsuccessful and the event is overlooked. This probability depends on k, the number of event packets required at the sink node to accurately characterize the event and the event detection radius r. Results from Fig. 5.5 show that less than 9 percent of the events are overlooked by selecting $k = 3$ and $r \geq 15$ m; we consider this to be adequate so $k = 3$ is used throughout this section. Nevertheless, in this section we also explore the rest of possible values of r depicted in Table 5.1.

Table 5.1: Network parameters.

Parameter	Value
Area Size	$100\,\text{m} \times 100\,\text{m}$
Sink node location	$(200, 0)$
Number of nodes in the area	$m = 100$
Data packet length	$l = 2$ kbits
Control packet length	$l_c = 1$ kbits
Data rate	$R = 40$ kbps
Time slot duration	$t_s = 0.1$ s
Energy consumed by the communication circuits	$E_{\text{elec}} = 50\,\text{nJ/bit}$
Low-power transmission range	$d_l = 35$ m
High-power transmission range	$d_h = \sqrt{200^2 + 100^2}$ m
Path loss exponent	$p_l = 2$
Energy consumed by the amplifier	$\epsilon_{\text{amp}} = 10\,\text{pJ/bit/m}$
Event detection radius	$r \in \{5, 10, 15, 20, 25, 30\}$ m
Number of event reports required at the sink	$k = 3$
Transmission probabilities	$\tau \in \{1, 2, \ldots, 100\} \cdot 10^{-2}$
Factor of reduction for τ	$b \in \{1, 2, \ldots, 10\}$

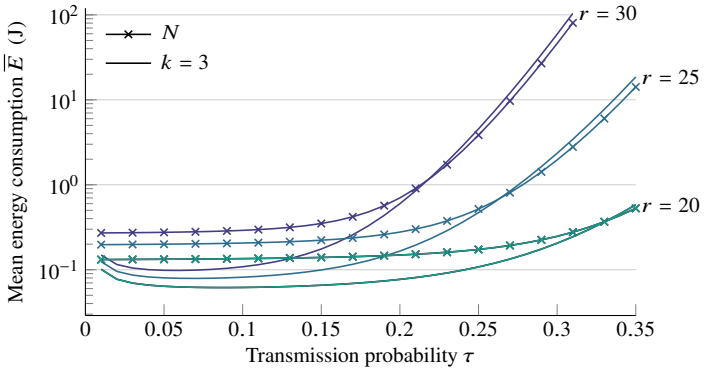


Figure 5.8: Mean energy consumption $\overline{E}$ for the transmission of $k = N$ and $k = 3$ event packets for the three longest detection radii $r \in \{20, 25, 30\}$ m, and $\tau \leq 0.35$.

5.4.1 FB approach

We begin our analysis of energy consumption in RA event reporting by comparing two different approaches. In the first one, each of the N detecting nodes is set to transmit a packet to the CH; this is the typical approach followed in the literature. In the second, CMs overhear the CH transmissions in order to identify the kth successfully transmitted packet. Then, the remaining packets are discarded to avoid energy wastage due to the transmission of redundant packets. Fig. 5.8 shows the mean energy consumption during event reporting $\overline{E}$ for both approaches. It is clear that energy consumption is greatly affected when high values of τ are selected. A similar but lesser effect is observed as the detection radius r increases. In other words, selecting lower values of τ and r is beneficial for energy efficiency, but the event overlooking probability increases as r decreases.

Furthermore, it can be seen from Fig. 5.8 that restricting the number of transmitted packets with overhearing reduces energy consumption for relatively low values of τ. In these cases, the network is not highly congested. Conversely, if high values of τ are selected, restricting the number of transmitted packets slightly increases the energy consumption with respect to the traditional approach. This is mainly caused by the congestion of the wireless medium.

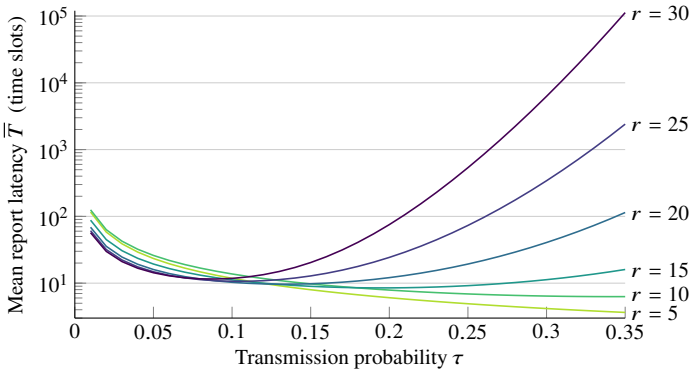


Figure 5.9: Mean report latency $\overline{T}$ assuming $k = 3$ event packets must be received at the sink node for several event detection radii r and transmission probabilities $\tau \leq 0.35$.

Next, we begin our analysis of report latency by showing its mean value in Fig. 5.9; hereafter we assume $k = 3$. Here we observe that high values of τ reduce report latency for small detection radii. On the other hand, high values of τ increase report latency for large detection radii. Hence, for large detection radii, report latency is correlated to energy consumption. This is relevant because selecting a large detection radius reduces event overlooking probability. Therefore, these results suggest the selection of a sufficiently long r and low τ is the most efficient solution for time-critical applications. Specifically, this approach sharply reduces report latency and event overlooking probability, while maintaining an adequate energy efficiency.

As stated earlier, obtaining the probability distribution of report latency provides with much more valuable information regarding the behavior of the system than mean report latency; this is especially important in time-critical applications. However, it may be difficult to select a proper metric to assess the report latency from its whole distribution. For the sake of simplicity, hereafter we assess the report latency in terms of its 90th percentile defined as

$$T_{90} \equiv \min_{s} \{s \mid F_T(s) \geq 0.9\}. \tag{5.31}$$

That is, 90 percent of the events are successfully reported in s or less time slots. Since the mean report latency and event overlooking probability are enhanced for large values

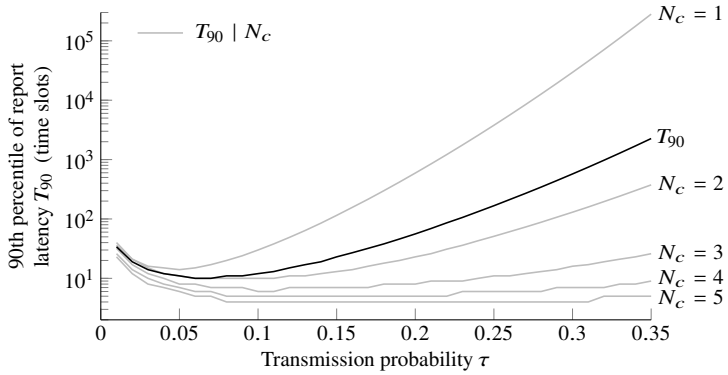


Figure 5.10: 90th percentile of report latency T_{90} for $r = 30$ m.

of r, we obtained the T_{90} for $r = 30$ m. Specifically, Fig. 5.10 shows the T_{90} given $N_c \in \{1, 2, 3, 4, 5\}$ and also T_{90} alone for the given scenario. Here, two tendencies are clearly observed: report latency steeply increases with τ and, the more clusters are involved in event reporting, the lower the time needed for receiving the required packets at the sink. Both of these tendencies confirm that the combination of low transmission probabilities with a large detection radius enhances event reporting. for instance, the minimum T_{90} is achieved by selecting $\tau = 0.06$. It is also important to observe that the report latency given $N_c = 1$ highly contributes to the overall report latency, whereas the contribution to this parameter fades considerably as N_c increases.

5.4.2 AB approach

In this subsection we use the hybrid method to analyze and optimize the performance of RA event reporting when an AB is implemented. As described previously, we first use our simulator to obtain the probability distribution of the number of detecting nodes, which allows us to calculate the energy consumption and report latency during event reporting analytically.

Our first step is to evaluate the energy efficiency of the AB. For this, the mean energy

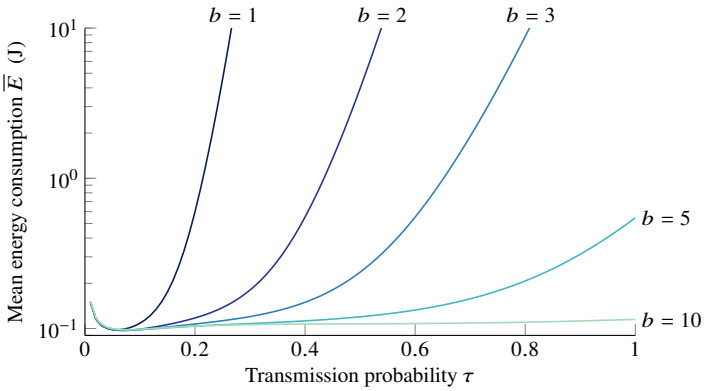


Figure 5.11: Mean energy consumption during event reporting $\overline{E}$ for the AB with $r = 30$ m.

consumption during event reporting given $r = 30$ m, $\tau \in (0, 1)$, and $b \in \{1, 2, 3, 5, 10\}$ is obtained and shown in Fig. 5.11. Please recall that $b = 1$ corresponds to the FB.

Clearly, $\overline{E}$ is a concave function, regardless of the value of b. Building on this, we define $\overline{E}^*(b)$ as the global minimum of $\overline{E} \mid b$ and $\tau^*(b)$ as the value of τ that leads to $\overline{E}^*(b)$; i.e., the only critical value of τ. Therefore, a value of τ that minimizes energy consumption can be obtained. It is clear from Fig. 5.11 that $\tau^*(b)$ lies around $\tau = 0.1$ for each of the selected values of b. Since τ only assumes discrete values with granularity 10^{-2}, $\tau^*(b)$ can be easily identified by simple search over nearby values of τ(i.e., brute force). The obtained values will be presented in Table 5.2.

A clear trend can be identified from Fig. 5.11: $\overline{E}$ grows rapidly with τ, given $\tau > \tau^*(b)$ and $b = 1$, but the rate of change is drastically reduced as b increases. On the other hand, $\overline{E}$ is greatly similar for every b when given $\tau < \tau^*(b)$. This is an intuitive result because collisions will rarely occur when τ is sufficiently small; hence, the transmission probabilities will rarely be modified. On the other hand, selecting a relatively high b reduces the negative impact of the selection of an exceedingly high value of τ when compared to $b = 1$. In other words, the robustness of the system increases with b.

We now proceed to investigate the behavior of report latency for each possible

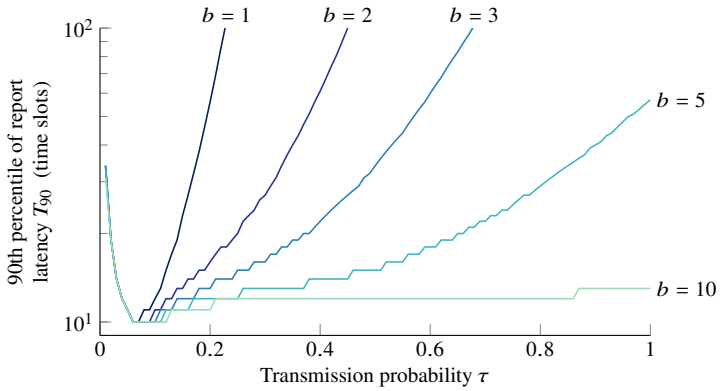


Figure 5.12: 90th percentile of the report latency T_{90} for the AB with $r = 30$ m.

value of $\tau \in (0, 1)$ and $b \in \{1, 2, 3, 5, 10\}$ given $r = 30$ m. The overall T_{90} obtained in this scenario is shown in Fig. 5.12. In this figure we observe a similar behavior than in Fig. 5.11, T_{90} is a concave function. However, the image of $\overline{E}$ is continuous whereas that of T_{90} is discrete. Consequently, numerous global minima and critical points can be identified in Fig. 5.12.

It is interesting to observe that $\tau^*(b)$, the critical points of $\overline{E} \mid b$, are also critical points for $T_{90} \mid b$, for any b. Therefore, we conclude that $\tau^*(b)$ optimize the performance of the RA protocol for a given b. The obtained critical points $\tau^*(b)$ are listed in Table 5.2, along with the achieved $\overline{E}$ and T_{90}.

Table 5.2 shows that $\tau^*(2) = 0.07$ leads to the global minimum energy consumption in the network. However, extremely similar values are obtained when selecting other values of b. Therefore, it is safe to say that an optimal performance can only be obtained with the AB. Still, a comparable performance can be obtained with the optimal configuration of the simple FB.

In Fig. 5.11 and Fig. 5.12 we observed that the robustness of the performance increases with b. In other words, increasing b widens the set of possible values of τ that lead to a near-optimal performance. At least in a scenario in which all events have similar characteristics (i.e., event detection radii r). We emphasize this latter statement by showing the relative increase in the 90th percentile in $\overline{E}$ and T_{90} due

Table 5.2: Transmission probabilities $\tau^*(b)$ that optimize performance for the given $b \in \{1, 2, 3, 5, 10\}$ and achieved $\overline{E}$ and T_{90}.

b	$\tau^*(b)$	$\overline{E}^*(b)$ (J)	$T^*_{90}(b)$
1	0.06	0.09812	10
2	0.07	0.09707	10
3	0.07	0.09714	10
5	0.07	0.09749	10
10	0.08	0.09838	10

to slight deviations from $\tau^*(b)$ for $b \in \{1, 2, 10\}$ in Fig. 5.13. The relative increase is calculated for each point as the ratio between the achieved QoS parameter and its global minima. It is observed that failing to select $\tau^*(1)$ highly affects performance. This is clear even for errors as low as $\tau = \tau^*(1) \pm 0.03$, for which an increase of up to 7 percent in$\overline{E}$ and up to 40 percent in T_{90} can occur with respect to the optimal $\tau^*(1)$. On the other hand, the energy consumption and report latency that the AB provides are much more robust to the inaccurate selection of parameters.

So far we have merely assessed the event report latency T in terms of its mean and 90th percentile. Results presented in Table 5.2 suggest there is no difference in this parameter regardless of the selected value of b, but this conclusion might not be precise. Hence, we now provide an in-depth look at the behavior of RV T by showing its complementary CDF (CCDF) (i.e., $1 - F_T(s)$) for $b \in \{1, 2, 10\}$ in Fig. 5.14. In other words, Fig. 5.14 shows the probability that event reporting has not been successful s time slots after event detection. Therefore, the lower the amplitude of the curve, the better the performance. Fig. 5.14 clearly shows that selecting $\tau^*(b)$ leads to an almost identical distribution of T and confirms our previous conclusion: a comparable performance can be obtained with the simple FB and AB approaches. Nevertheless, selecting $\tau^*(2)$ will lead to the minimum report latency in most occasions. The opposite occurs when selecting $\tau^*(10)$.

This concludes our analysis in simple single-event environments. Hence, the

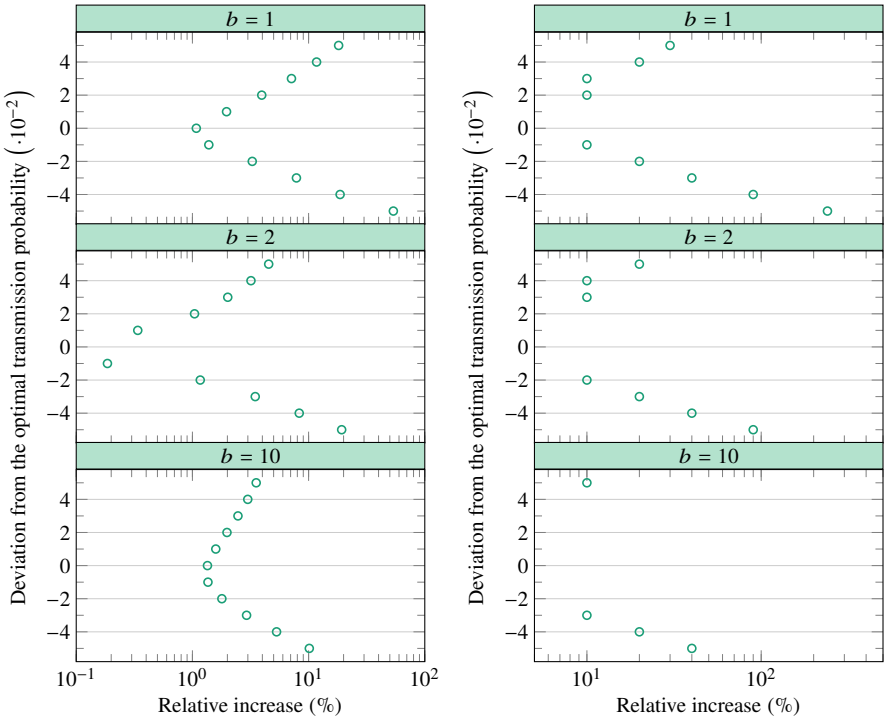


Figure 5.13: Relative increase in the (a) mean energy consumption $\overline{E}$ and (b) 90th percentile of report latency T_{90} due to slight deviations from $\tau^*(b)$ given $r = 30$.

following subsection focuses on the performance of RA in slightly more complex multi-event environments.

5.4.3 Multi-event environments

In multi-event environments, the network is in charge of monitoring several types of events and each of them presents different characteristics. This is a typical scenario in complex WSN applications as nodes may include a wide arrange of sensors for

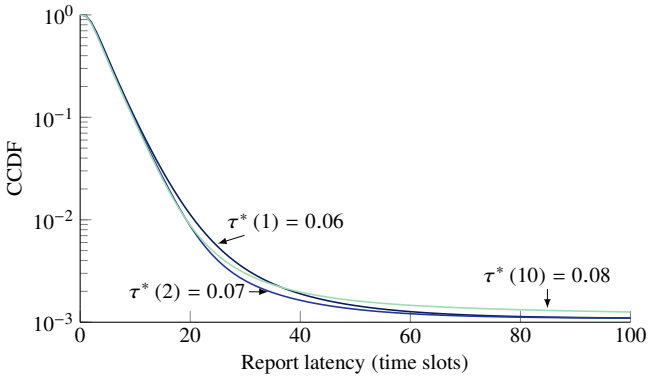


Figure 5.14: CCDF of report latency $1 - F_T(s)$ given $r = 30$ m.

different physical parameters.

The remainder of this section is dedicated to evaluate and optimize the performance of RA event reporting in an environment where two types of events occur. Specifically, we assume an environment in which the detection radius of 75 percent of the events is $r = 30$ m and the detection radius of the remaining 25 percent of the events is $r = 15$ m. That is, $\Pr[r = 30] = 0.75$ and $\Pr[r = 15] = 0.25$.

We used our hybrid method to obtain $\tau^*(b)$ for each b, and the combination of $\tau^*(b)$ and b as described in the previous subsection. Table 5.3 shows $\tau^*(b)$ and the achieved QoS parameters. This table reveals that, again, selecting $\tau^*(2)$ leads to the global minimum energy consumption, but also to the same T_{90} with any b. Yet another interesting result is that the values $\tau^*(b)$ in the multi-event environment are greatly similar to those obtained in the single-event environment. This result suggests that there is no need to possess a deep knowledge on the characteristics of the events (i.e., event detection radii) to achieve a near-optimal performance. We further investigate the validity of this latter statement by obtaining the relative increase in the QoS parameters due to inaccurate selection of τ.

Fig. 5.15 shows the relative increase in $\overline{E}$ and T_{90} due to slight deviations from $\tau^*(b)$. Here we observe that the implementation of an AB in multi-event environments leads to a similar behavior as in single-event environments. That is, $b = 2$ minimizes

Table 5.3: Transmission probabilities $\tau^*(b)$ that optimize performance for the given $b \in \{1, 2, 3, 5, 10\}$ and achieved $\overline{E}$ and T_{90} in the multi-event environment.

B	$\tau^*(b)$	$\overline{E}^*(b)$ (J)	$T^*_{90}(b)$
1	0.07	0.08456	13
2	0.07	0.08287	13
3	0.07	0.08293	13
5	0.08	0.08321	13
10	0.08	0.08406	13

$\overline{E}$ and increasing, and leads to the second and third lowest $\overline{E}$ even if slight errors are made in the selection of τ. Furthermore, increasing b increases the robustness of event reporting in the sense that the stability of the achieved QoS parameters increases with b. However, Table 5.2 and Table 5.3 show that the energy efficiency drops slightly as b increases. Hence, selecting an arbitrarily large value of b is not recommended.

We conclude our analysis of event reporting in multi-event environments by showing the CCDF of RV T within each independent cluster for $b \in \{1, 2, 10\}$ given $\tau^*(b)$ in Fig. 5.16. In other words, Fig. 5.16 shows the probability that event reporting has not been successful in one cluster s time slots after event detection. As for Fig.5.14, the lower the amplitude of the curve, the better the performance. Fig. 5.16 illustrates the same behavior as Fig. 5.14: selecting $\tau^*(2)$ will lead to the shortest report latency in most occasions. On the other hand, selecting $\tau^*(10)$ will lead to the highest report latency (when compared to $b \in \{1, 2\}$) in most occasions. However, here the differences between the selected $\tau^*(b)$ are much more noticeable than in Fig. 5.14. The reason for this is that selecting $b = 10$ when events with a relatively short event radius occur (i.e., $r = 15$), leads to an excessively low β after a collision occurs. This in turn causes a relatively large number of time slots in which no BCM attempts transmission.

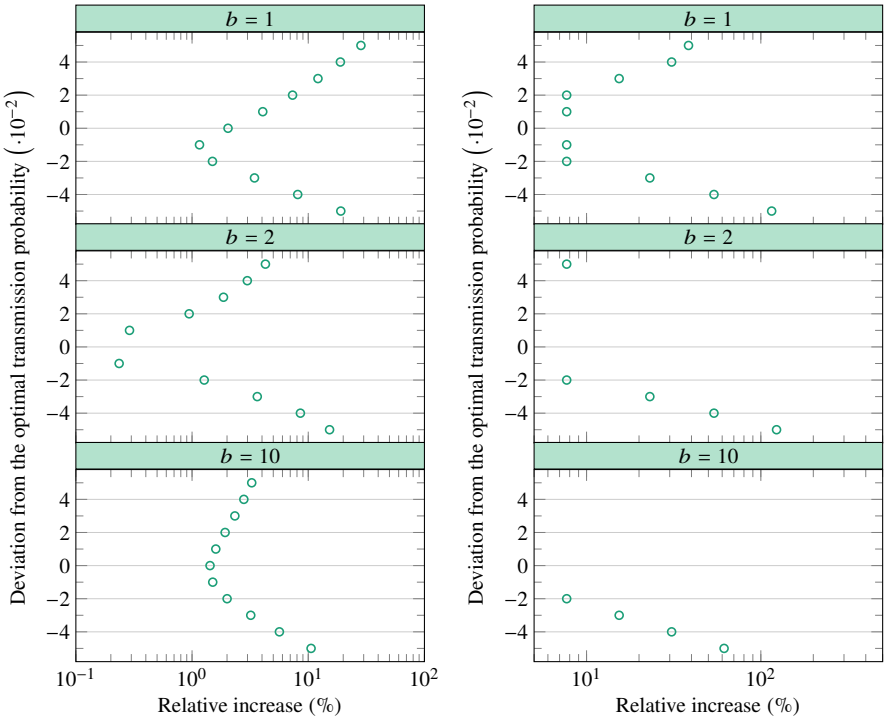


Figure 5.15: Relative increase in the (a) mean energy consumption $\overline{E}$ and (b) 90th percentile of report latency T_{90} due to slight deviations from $\tau^*(b)$ in the multi-event environment.

5.5 Conclusions

This chapter presented a hybrid method for the QoS analysis of RA WSN protocols that is capable of obtaining the pmf of report latency. Because of this, our method is especially useful to assess the QoS of WSNs in time-critical applications, where event reporting is time-constrained and fault-sensitive. It also considers a basic structure and can be easily adapted to accommodate a wide range of routing protocols and to

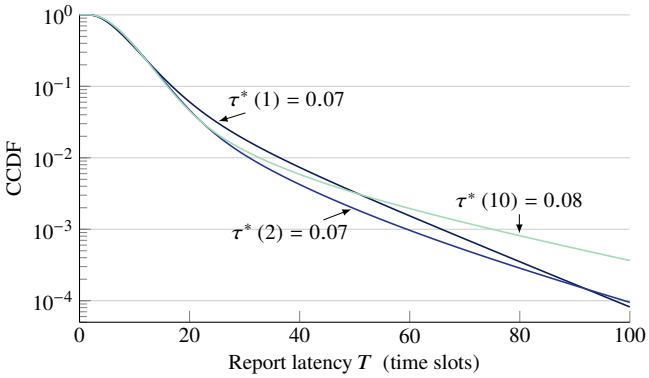


Figure 5.16: CCDF of report latency within one cluster given $\Pr[r = 15] = 0.25$, $\Pr[r = 30] = 0.75$.

analyze a wide range of MAC protocols.

In our method, the pmf of detecting nodes is obtained by simulation and allows the calculation of the overlooking probability, given the event detection radii and the required number of messages to characterize the occurring event are known. Then, we use DTMCs to obtain the two most important QoS parameters in WSNs: energy consumption and report latency.

Our simulation results confirmed an intuitive trade-off: large detection radii reduce the event overlooking probability but may increase congestion and, hence, energy consumption during event reporting. However, an adequate selection of the parameters of the RA protocol can effectively relieve congestion, but also minimize energy consumption and report latency. This latter statement was emphasized in Section 5.4, dedicated to the performance analysis and optimization of a simple RA protocol. That is, results presented showcase the importance of the adequate selection of transmission probabilities for the given event detection radius.

It is clear that an optimal performance can only be obtained with the optimal configuration. Nevertheless, this implies having perfect information on the characteristics of the occurring phenomena. In our case, this is represented by knowing the exact detection radii of the events. Still, results obtained with our hybrid model suggest that,

even if the knowledge of the occurring phenomena is scarce, a near-optimal performance can be obtained by following some recommendations: 1) set a relatively low threshold in the nodes to achieve a relatively long detection radius; 2) implement a RA protocol with overhearing and set each cluster to transmit three event reports to eliminate redundant transmissions; 3) set the transmission probabilities to a relatively low value, for example 0.1; and 4) implement an AB approach with $b = 2$, this mitigates the negative effects of the inaccurate selection of the transmission probability. These recommendations hold under both, single-event and multi-event environments.

Chapter 6

Network-coded cooperation (NCC) for efficient massive content delivery through cellular networks

6.1 Introduction

Wireless data traffic is increasing dramatically. For instance, the amount of traffic transmitted in 2016 grew 63 percent when compared to 2015. By 2021, a data traffic of 49 exabytes per month by 2021 is expected. This represents an increase of around 700 percent and from 60 to 78 percent of this traffic will be caused by mobile video. Furthermore, cellular data consumption is expected to rise from slightly less than 1 GB in 2016 to around 5.7 GB per month in 2021; that is more than a five-fold increase in five years [32]. Such a dramatic increase in data traffic poses important challenges to the actual 4th generation (4G) technology that are completely different to those investigated Chapters 2 to 5, where massive machine-type communication (mMTC) was investigated. Instead, in this chapter we focus on the second of the three main use cases for 5th generation (5G): enhanced mobile broadband (eMBB).

Since the commercialization of 2nd generation (2G) in 1991, where digital wireless communications were first introduced to the phone industry, the evolution of mobile

technology has focused on merely increasing data rate. That is, the achievable data rate in mobile phones has increased from less than 1 Mbps in 2G, to around 60 Mbps in 3rd generation (3G), and to up to 500 Mbps in 4G. Nevertheless, few advances have been made in other aspects such as latency and, more importantly, the number of users supported by the mobile base stations. These are key to provide efficient gaming, virtual/augmented reality, and video streaming applications to the masses. Hence, are key components of the 5G eMBB.

For instance, the user equipments (UEs) that request access to a given content (e.g., video streaming) through LTE Advanced (LTE-A) are connected via a unicast link from the cellular base station (evolved NodeB (eNB)). This is true even though the exact same content is transmitted to these UEs simultaneously. Therefore, a large number of replicated unicast sessions are created in the latter case. For example, such a scenario can occur with the passengers in the same train, with the audience in a music festival or in a football stadium, or by players in augmented reality mobile games.

The industry is aware that the current LTE-A system will not be able to handle the expected increase in data traffic in the coming years. Consequently, several systems have been deployed in order to provide multicast capabilities to LTE-A. A system that took advantage of this necessity in the early 2010s was the LTE-A evolved multimedia broadcast multicast service (eMBMS) [104]; a multicast implementation through LTE-A small cells. However, several issues were detected during its implementation. For example, the eMBMS may suffer from unexpected disconnections [26], reduced transmission range, high energy consumption, and poor spectral efficiency [36]. Therefore, different content delivery mechanisms to reduce the amount of traffic requested directly from cellular networks must be designed.

Cooperative mobile clouds (MCs) are a promising solution to the described content delivery scenario [82]. An MC is a cooperative architecture in which a group of UEs share the available wireless resources opportunistically. For instance, UEs cooperate through a short-range technology, such as WiFi, which sharply reduces the consumed resources in the LTE-A link [39]. Hence, MCs may drastically offload the data traffic at the eNB, but can also provide many other benefits; some of these benefits will be investigated throughout this chapter. It is important to mention that WiFi and

cellular interfaces are not entirely integrated in 4G. On the other hand, 5G promises full integration with these two interfaces [19]. This greatly increases the potential benefits of MCs in 5G.

Some content delivery systems that combine long and short-range technologies have been proposed in the literature, but short-range unicast sessions are oftentimes used [22, 54]. Since the UEs within an MC are closely located, the use of multicast short-range links for content delivery is possible and much more efficient than the use of independent unicast sessions. It is in multicast wireless networks where novel approaches, such as network coding (NC). have proven to be highly valuable to ensure a high data rate and a low error rate [15].

In traditional data transmission, data packets are created by joining a header (i.e., control information) and the payload; this (source) packet is transmitted and, hopefully, it will reach the destination. Depending on the technology, different feedback mechanisms can be used. Hence, a lost packet triggers a retransmission request from the destination. For example, hybrid ARQ (HARQ) is used in LTE-A, where redundancy is added to each retransmission until it is received without errors.

Instead, NC is a novel communications paradigm in which linear combinations of packets are created; these are transmitted instead of the original source packets. For this, the transmitter combines a batch of packets contained in its coding matrix, known as a generation, to produce coded packets. As such, the receiver only needs to receive sufficient linearly independent packets to decode the whole batch [15].

Random linear NC (RLNC) is one of the most widely used NC schemes [47]. In RLNC, each packet in the generation is multiplied by a coefficient chosen randomly from a Galois-field of size q, denoted as GF(q). Then, these packets are combined to form a coded packet; the coded packet is sent along with the coding coefficients so the receiver can decode the batch. The systematic RLNC is a variant of RLNC in which the source packets are first sent one after the other. Then, coded packets are transmitted to recover the errors that may have occurred during the transmission of the source packets. Fig. 6.1 presents an example to transmit three data packets with traditional feedback mechanisms, full-vector RLNC, and systematic RLNC. In this particular example, packets are transmitted from right to left, and the second and

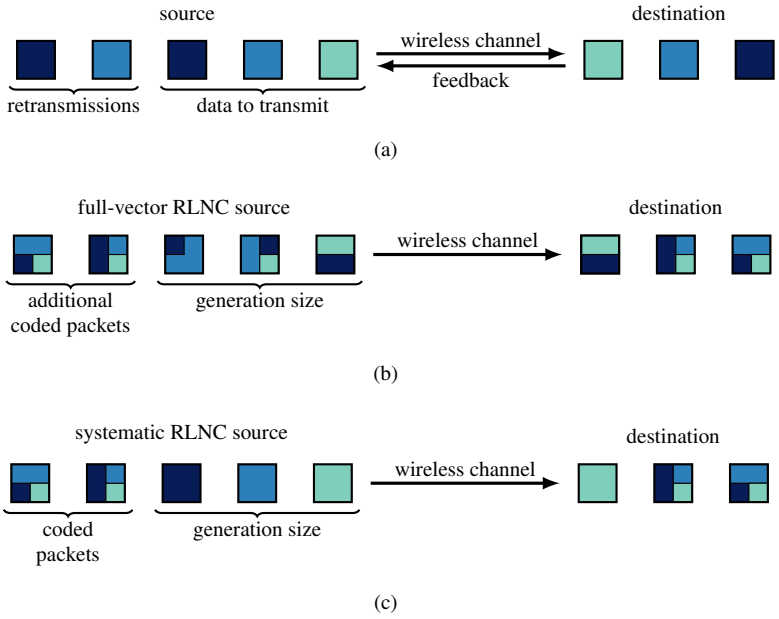


Figure 6.1: Example to transmit three data packets with: (a) traditional feedback mechanisms; (b) full-vector RLNC; and (c) systematic RLNC.

third transmissions are lost due to wireless channel errors. As it can be seen, RLNC schemes provide the valuable benefit of eliminating feedback messages.

In addition, research has shown that less packet transmissions are needed with the systematic RLNC than with the traditional full-vector RLNC (i.e., in which only coded packets are transmitted). This considerably reduces energy consumption and increases throughput. As an added benefit, systematic RLNC reduces the decoding complexity at the UEs when compared to full-vector RLNC [52]. It is clear from Fig. 6.1 that this difference in the computational complexity is due to the fact that less packets have to be coded and decoded. Therefore, it occurs in both, the source and the destination.

The combination of cooperative approaches such as MCs with RLNC schemes has lead to the innovative communication paradigm of NCC [55, 87]. NCC has the potential to provide increased performance in multicast applications when compared to

either MCs and RLNC alone [47]. In this chapter, we propose a simple NCC protocol for the efficient content delivery in cellular networks. It comprises two phases, namely the cellular and MC phases. In the cellular phase, the eNB segments the requested content in batches of size g packets; hereafter we refer to the batch size g as the generation size. These g packets are transmitted to an MC through time multiplexed unicast links. Then, in the MC phase, the UEs cooperate under a distributed systematic RLNC scheme to distribute these packets through multicast WiFi links.

One of the main drawbacks of existing cooperative systems is the transmission of a large number of feedback messages within the MCs, which are needed to keep track of the state of the UEs [27]. Hence, in this study we eliminate the transmission of feedback messages from the UEs and instead use a simple but accurate analytical model to calculate and minimize the number of coded packet transmissions needed to achieve a predefined reliability of distributing the whole generation to the MC. This approach solves the problem of excessive feedback and also optimizes the utilization of resources.

However, two main challenges arise when modeling the multicast transmissions under an RLNC scheme. The first challenge is to model a multicast problem with multiple sources. That is, the content is distributed among the UEs in the MC and the packets received at each node are not present at the remaining UEs. Single-source multicast scenarios under RLNC schemes have been studied in the literature and the formulation of the exact decoding probability is complicated [101]. Concretely, exact formulations only exist for the case of one source and two destinations [55] and lower bounds must be used for a higher number of destinations. In our model, we incorporate a lower bound to solve this problem, whose accuracy under a simple multicast setup that incorporates the systematic RLNC scheme has been confirmed in [101].

The second challenge is to model the inclusion of packets received from both, the eNB and MC neighbors in the coding matrix of the UEs. This approach enhances the throughput when compared to other policies like, for example, only include packets received directly from the eNB in the coding matrix [99]. Needless to say, finding an accurate expression for the linear independence of every coded packet transmission in our scenario is also a cumbersome task. Therefore, we use an upper bound for the

probability of linear independence of coded packets and evaluate its accuracy; it is described in Section 6.4.

We also use our analytical model to calculate the energy consumption at the UEs. Our results show that energy savings of more than 37 percent when compared to single unicast content delivery can be achieved with our protocol in addition to a reduced LTE-A bandwidth utilization. Specifically, the exact same amount of LTE-A resources utilized by a single-user unicast download are needed for each MC. Hence, LTE-A bandwidth gains grow linearly with the number of UEs in the MC.

The rest of the chapter is organized as follows. Section 6.2 presents the state of the art of cooperative systems for massive content delivery. Then, Section 6.3 presents the proposed NCC protocol and Section 6.4 presents the formulated analytical model, including the process for the optimization of our NCC protocol. Section 6.5 presents our main results and their implications. Finally, Section 6.6 presents our conclusions.

6.2 Related work

As mentioned above, the eMBMS system has several drawbacks. One of the most important ones is that it suffers from unexpected disconnections and lacks mechanisms to provide quality of service (QoS) guarantees. Therefore, diverse solutions to the massive content delivery in LTE-A have been developed. For instance, the idea of organizing microcells in cloudlets was first described in [38]. Cooperative relaying was proved to increase network performance in [56] whereas Ahlswede *et.al* advocated the concept of network information flow and its advantages in [15]; this work is considered the precursor of NC schemes. Moreover, the interplay between subgrouping in cloudlets and NC was first proposed in [87].

Despite the clear advantage of short-range NC multicast in the cloud, most existing cooperative systems consider unicast short-range data transmissions. Some examples are the MicroCast [54], and CoopStream [22] systems, whose main focus is to offload data traffic from the eNB. Clearly, the performance of all these previous technologies might increase by using WiFi multicast in the short-range.

Wang *et al.* presented a new NC-based video conference system for mobile devices in multicast network called NCVCS [105]. NCVCS demonstrates the advantages of multicast over short-range unicast, but lacks the cellular communication backhaul.

The main motivation for this work is the NCC system first proposed in [100]. The main focus of this NCC network was to offload the LTE-A network, but also, important throughput and energy gains were observed. Consequently, demonstrators were built and presented at MWC 2017, and IEEE CCNC 2018/CES 2018 [80], and IEEE 5G Summit 2018. This latter demo was the product of my collaboration with the Deutsche Telekom Chair of Communication Networks of the Technische Universität Dresden, Germany, during a six-month research stay.

Regarding the analytical modeling of RLNC multicast, a thorough study on the decoding probability in a one-source multicast scenario with both, full-vector and systematic RLNC was conducted by Tsimbalo *et al.* [101]. Specifically, Tsimbalo *et al.* define the probability of successful content delivery to be the probability that every node receives the whole generation. Please observe that this definition magnifies the importance of considering the correlation between the packets received at each node to calculate the desired probability. Tsimbalo *et al.* concluded that the effect of this correlation may be negligible only if the systematic RLNC is used. On the other hand, neglecting the effect of correlation can affect the accuracy of models if full-vector RLNC is used. As it will be seen in Section 6.4, we deal with a similar but even more complex problem because in our NCC protocol: a) content distribution within the MCs is performed through multiple multicast sessions; b) the eNB distributes the data packets among the UEs; and c) coding is performed by combining the packets received from the eNB and from neighboring UEs. Therefore, we propose a different definition for the probability of successful content delivery that increases the accuracy and simplicity of our calculations when compared to the definition provided by Tsimbalo *et al.* [101].

Table 6.1: Comparison between related systems

		LTE-A	NC	Short-range
eMBMS	[26]	✓	✗	✗
MicroCast	[54]	✓	✓	Unicast
CoopStream	[22]	✓	✓	Unicast
NCVCS	[105]	✗	✓	Multicast
NCC system	[100]	✓	✓	Multicast

6.3 NCC protocol and basic assumptions

In this section we describe the NCC protocol we propose for massive content delivery through cellular networks. In addition, we outline simple assumptions used for our analysis.

As a starting point, groups of UEs called MCs are formed. For this, let n be the maximum number of UEs that are allowed in an MC, hereafter referred to as the cloud size; n is signaled by the eNB as a configuration parameter. MCs are groups of at most n UEs that: a) have LTE-A connection to the same eNB; b) request access to the exact same content; and c) are fully interconnected by a short-range technology, namely WiFi. It is out of the scope of our study to develop the rules and the protocol for the formation of MCs. Instead, we focus on the content delivery once the MCs have been formed. Nevertheless, a similar approach to that of clustering algorithms for wireless sensor networks (WSNs), such as the one described in Chapter 5, can be used.

Content delivery occurs in two phases: the cellular and MC phases. In the cellular phase, the eNB segments the requested content in batches of g data packets; g is commonly known as the generation size. These g packets are transmitted to an MC through n unicast sessions. Then, at the MC phase, the UEs first multicast the packets received from the eNB without coding. Afterwards, the UEs multicast coded packets to recover the errors that may have occurred in the previous transmissions. Fig. 6.4 illustrates the basic operation of our NCC protocol in the cellular and MC phases;

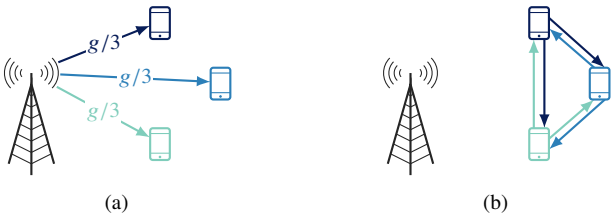


Figure 6.2: Overview of the (a) cellular and (b) MC phases that comprise our NCC protocol.

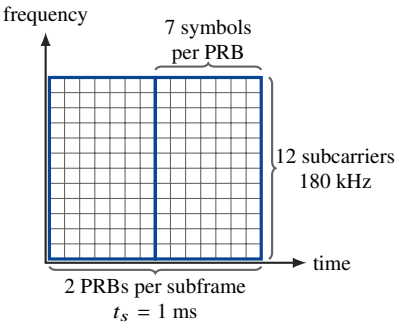


Figure 6.3: Structure of the physical resource blocks (PRBs) in LTE-A.

these are described in detail in the following.

Cellular phase: The eNB transmits the g data packets to the n UEs through n unicast sessions. In LTE-A, data transmission takes place in a slotted channel, whose minimum scheduling unit is one subframe, with duration $t_s = 1$ ms [8]. The minimum unit for data downlink transmission in LTE-A is the physical resource block (PRB), which is defined as the set of 7 of consecutive orthogonal frequency-division multiplexing (OFDM) symbols in the time domain and 12 consecutive subcarriers in the frequency domain [8]; in the time domain, two PRBs fit in one subframe. Fig. 6.3 illustrates the structure of the PRBs.

We assume the n unicast sessions are multiplexed, either in time or in frequency, so packets are transmitted to the UEs in a round-robin fashion. For this, each of the

n UEs is assigned an index, in the set $\mathcal{N} = \{i \in \mathbb{Z}_+ \mid i \leq n\}$, that defines the order in which they will receive the data packets from the eNB. Therefore, if time-division multiplexing (TDM) is used, only one data packet transmission occurs simultaneously at each cloud. On the other hand, if frequency-division multiplexing (FDM) is used, the number of simultaneous data packet transmissions at each MC is the minimum between n and the maximum number of simultaneous data packet transmissions that can be accommodated in one cellular carrier. The latter is determined by the cell bandwidth and the selected data rate. Please observe FDM unicast may not be possible under certain applications that generate the data on the fly. Live video streaming applications are clear examples of such applications. For these, TDM unicast must be used.

Throughout this chapter, we assume the MCs are closely located to the eNB so that no wireless channel errors can occur during the cellular phase. This is a valid assumption as the considered data rate is relatively low (see Table 6.3 on page 175), LTE-A is provided with highly reliable data transmission mechanisms such as HARQ, and is set to modify the modulation and coding scheme (MCS) if the packet erasure ratio (PER) is higher than 0.1 [4, Sec. 7.2.3]. As such, the cellular phase is comprised of g transmissions, distributed among the n UEs, which must cooperate to distribute these packets in the MC phase.

MC phase: The UEs are in charge of redistributing the g packets received from the eNB through the MC. No feedback messages are transmitted in this phase, so the eNB must inform the number of time slots allocated for the content distribution within the MC to the UEs.

The index i assigned to each UE in the cellular phase is used to create a time-division multiple access (TDMA) schedule. At each time slot, a UE performs a WiFi multicast packet transmission to the remaining $n - 1$ UEs in the MC. Therefore, the set of neighbors of the ith UE is $\mathcal{N}_i = \{j \mid j \in \mathcal{N} \setminus i\}$ and has $n - 1$ elements. The transmitting UE changes at each time slot to uniformly distribute energy consumption among the MC members. Please observe that the time slot duration at this phase is not necessarily the same as that of the LTE-A subframe, hence a higher or lower data rate can be used.

At the end of the cellular phase, g_i packets are present at the ith UE and these are not present in the remaining $n-1$ UEs. A distributed systematic RLNC scheme is implemented in this phase, hence, g_i source (i.e., not coded) packets are transmitted by the ith UE. In other words, UEs relay the source packets received from the eNB, but not from their neighbors. Therefore, the first g packet transmissions within the MC are not coded. Then, coded packet transmissions are performed in order to recover the errors that occurred during the g source packet transmissions.

Exactly g time slots are needed for the transmission of the g source packets. Therefore, the eNB only has to calculate the number of time slots allocated for the transmission of coded packets s. An MC phase concludes when $g+s$ time slots have elapsed. Then, the eNB continues with the transmission of the next generation if needed, hence a new cellular phase begins. Otherwise, data transmission is terminated.

The timing diagram at each phase of our NCC protocol for $n=3$, $g=5$, and $s=2$ is illustrated in Fig. 6.4. Naturally, the same amount of resources are utilized when either TDM or FDM are used in the cellular phase. But FDM greatly reduces the cellular phase length. In the MC phase depicted in the diagram (right block), wireless channel errors occur at the second and fourth time slots. Each of these errors is recovered with one of the two coded packet transmission because UEs include packets transmitted by neighboring UEs in their coding matrices.

It is important to mention at this point that cellular and MC can be performed in sequence or in parallel, depending on the multi connectivity capabilities of the UEs (i.e., smart phones) in the MC. That is, if cellular and WiFi interfaces can be used simultaneously, these phases can occur in parallel after the first source packet is transmitted by the eNB. Otherwise, these must be performed in sequence. Current 4G smart phones are not likely to support this kind of multi connectivity. On the other hand, 5G promises full integration between these two technologies, hence, it is only natural to assume this kind of multi connectivity will be widely available in commercial devices. Hereafter, we refer to the case in which cellular and MC phases are performed in sequence as the 4G scenario. Conversely, we refer to the case in which cellular and MC phases are performed in parallel as the 5G scenario.

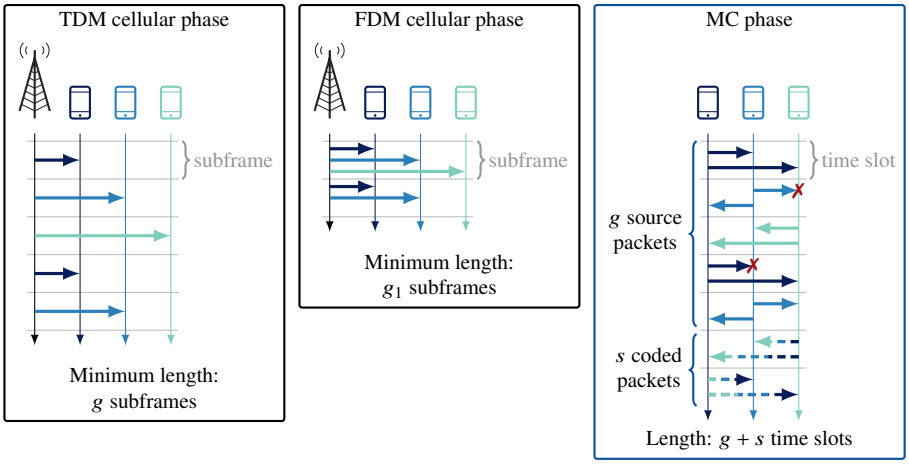


Figure 6.4: Timing diagram for the proposed NCC protocol given $n = 3$, $g = 5$, and $s = 2$. The errors that occurred at the second and fourth time slots are recovered with the coded packet transmissions.

6.4 Analytical model

In this section we provide a simple but accurate analytical model to optimize our NCC protocol. For this, let s be the number of coded packet transmissions performed in the MC. That is, from every $i \in \mathcal{N}$; please observe s is selected by the eNB and informed to the UEs. Building on this, we seek to obtain s^*, defined as the minimum value of s needed to achieve a desired reliability τ. Once s^* has been obtained, the maximum throughput and the average energy consumption per UE can be easily calculated.

To find s^*, let S be the random variable (RV) that defines the total number of coded packet transmissions needed so the n nodes at the MC decode the generation. Therefore, S has a phase-type (PH) distribution that describes the probability that the coding matrices of the UEs in the MC are full rank. A coding matrix is full rank when it has exactly the same number of columns and linearly independent rows. The linearly independent rows are known as degrees of freedom (DOFs). The rank of $\mathbf{D}$ can be calculated by performing Gaussian elimination so that the matrix is in reduced row

$$\begin{bmatrix} 1 & X & X & X \\ 0 & 1 & X & X \\ 0 & 0 & 1 & X \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Figure 6.5: Example of a full-rank 4×4 matrix.

echelon form and counting the number of ones in the diagonal, these ones are known as pivots. Fig. 6.5 illustrates a 4×4 matrix that is full-rank.

Hereafter, we refer to S as the probability of successful content delivery, whose support is the number of time slots allocated for the transmission of coded packets s. Building on this, s^* is defined as

$$s^* \equiv \min_s \{s \mid F_S(s; n) \geq \tau\} \tag{6.1}$$

where F_S is the cumulative distribution function (CDF) of RV S. That is, τ is a threshold for S and its value must be selected depending on the needs of the content delivery application. The process to calculate S is described in the following.

At the end of the cellular phase, g source packets have been distributed among the n UEs in the MC following a round-robin scheduling. We define g_i as the total number of data packets received by the ith UE in the cellular phase. Clearly, g_i is also the number of source packets transmitted by the ith UE in the MC phase, and is given as

$$g_i = \left\lceil \frac{g - (i-1)}{n} \right\rceil. \tag{6.2}$$

To proceed, please recall the set of neighbors of the ith UE is $\mathcal{N}_i = \{j \mid j \in \mathcal{N} \setminus i\}$. From there, we define the time index s_i as the number of coded packet transmissions towards the ith UE (i.e., from every $j \in \mathcal{N}_i$). Also let ϵ be the PER at the WiFi links (i.e., we assume the same ϵ for each pair of UEs $\{i, j\}$). Building on this, we define the stochastic process $\{X_{s_i}(i)\}_{s_i \in \mathbb{N}}$ as the rank of the coding matrix of the ith UE at time index s_i, whose support for any s_i is $x = \{0, 1, \ldots, g\}$. The RV of the stochastic

process defined above at $s_i = 0$ $X_0(i)$ is our starting point and has special significance as it defines the rank of the coding matrix of the ith UE at the end of the source packet transmissions. The probability mass function (pmf) of $X_0(i)$ is

$$p_{X_0}(x;i) = \Pr[X_0(i) = x] = \binom{g - g_i}{x - g_i}(1-\epsilon)^{x-g_i}\,\epsilon^{g-x}. \tag{6.3}$$

Please observe that, since only source packets have been transmitted up to this point, $X_0(i)$ is also the number of non-zero columns in the coding matrix of the ith UE at $s_i = 0$.

Next, coded packet transmissions are performed at every $s_i \geq 1$. Let, $S(i)$ be the RV that defines the number of coded transmissions from the $j \in \mathcal{N}_i$ UEs needed for the coding matrix of the ith UE to be full rank. $S(i)$ also has a PH distribution whose domain is the set of values for time index s_i. We calculate the CDF of $S(i)$ as

$$F_S(s_i;i) = F_{X_{s_i}}(g;i) = \Pr\left[X_{s_i}(i) = g\right]. \tag{6.4}$$

Clearly, $S(i)$ depends on the PER, denoted as ϵ, and on the probability of linear independence of each of the s_ith coded packet transmissions, denoted as $P_{\text{li}}(s_i)$. Nevertheless, the correlation between the packets received at each pair of UEs is needed in order to obtain the exact value for $P_{\text{li}}(s_i)$. Therefore, we define the stochastic process $Z_{s_i}(i,j)$ as the number DOFs that are missing from the coding matrices of both, the ith and jth UEs at s_i. The joint pmf of $X_0(i)$ AND $Z_0(i,j)$ is given as

$$p_{X_0Z_0}(x,z \mid i,j) = \epsilon^{g-x+z}\sum_u \left[\binom{g_j}{u}\binom{\gamma}{x-g_i-u}\binom{\gamma-x+g_i+u}{z}(1-\epsilon)^{\gamma+u-z}\right] \tag{6.5}$$

where $\gamma = g - g_i - g_j$ and u represents the number of DOFs in the coding matrix of the ith UE that were transmitted by the jth UE. The summation in (6.5) is performed in the set of possible values

$$\left\{u \in \mathbb{N} \mid \max\{0, x-\gamma-g_i+z\} \leq u \leq \min\{g_j, x-g_i\}\right\}.$$

The exact value of $P_{\text{li}}(s_i)$ for a given x and z is defined as

$$\begin{aligned} P_{\text{li}}(s_i \mid x,z) &= \Pr\left[X_{s_i+1}(i) = x+1 \mid X_{s_i}(i) = x \cap Z_{s_i}(i,j) = z\right] \\ &= 1 - q^{x+z-g}. \end{aligned} \tag{6.6}$$

That is, $P_{\text{li}}(s_i)$ depends on x and z, but also on the selected Galois-field size q and on the generation size g.

Clearly, different pairs of UEs $\{i, j\}$ may have different joint distributions of $X_0(i)$ AND $Z_0(i, j)$, as these depend on g_i, g_j, and γ. Furthermore, the joint pmf of $X_{s_i}(i)$ AND $Z_{s_i}(i, j)$ is different for each s_i. That is, the joint pmf of $X_{s_i}(i)$ AND $Z_{s_i}(i, j)$ must be calculated for every possible s_i and for each $\{i, j\}$ in order to calculate the exact $P_{\text{li}}(s_i)$. This makes our problem intractable even for small values of n and s_i. For instance, a related problem has only been solved for one transmitter and two receivers by Khamfroush *et al.* [55], but no exact formulations exist for a higher number of receivers.

Instead, we approximate $P_{\text{li}}(s_i)$ by assuming that every one of the missing DOFs in the decoding matrix of the ith receiving UE is present in the coding matrix of the jth transmitting UE. That is, $\Pr\left[Z_{s_i}(i, j) = 0\right] = 1$ for each s_i, i, and j, which gives

$$P'_{\text{li}}(s_i) = P_{\text{li}}\left(s_i \mid x, 0\right) = 1 - q^{x-g}; \tag{6.7}$$

this is clearly an upper bound for $P_{\text{li}}(s_i)$ and allows us to use the pmf of $X_{s_i}(i)$ alone instead of the joint pmf of $X_{s_i}(i)$ AND $Z_{s_i}(i, j)$ to calculate $S(i)$.

Naturally, (6.7) is exact for for $q \approx \infty$ and also for $n = 2$ since $g = g_1 + g_2$ in this latter case. The mean squared error (MSE) of the upper bound in (6.7) can be calculated as

$$\text{MSE}\left[P'_{\text{li}}(s_i)\right] = \sum_{\forall x,z} p_{X_{s_i} Z_{s_i}}\left(x, z \mid i, j\right)\left(q^{x+z-g} - q^{x-g}\right)^2. \tag{6.8}$$

Table 6.2 shows the MSE for the first coded transmission in the MC, $\text{MSE}\left[P'_{\text{li}}(0)\right]$, for characteristic values of n, g, ϵ, and q. The first coded transmission for $n = 3$ and for $n = 100$ is performed by the second and the first UEs, respectively. Therefore, $\text{MSE}\left[P'_{\text{li}}(0)\right]$ was obtained with $i = 1$ and $j = 2$ for $n = 3$, and with $i = 2$ and $j = 1$ for $n = 100$.

Clearly, (6.7) provides a highly accurate approximation, and the parameter that has the greatest impact on accuracy is the field size q. Concretely, a relatively high error is only obtained with $q = 2$ by setting: a) a short g and high PER; and b) a large g and

Table 6.2: MSE between the approximate and exact probability of linear independence of the first coded packet transmission.

	$n = 3$		$n = 100$	
	$g = 10$	$g = 100$	$g = 10$	$g = 100$
$\epsilon = 0.02$				
$q = 2$	$6.95 \cdot 10^{-5}$	$3.09 \cdot 10^{-4}$	$1.77 \cdot 10^{-4}$	$5.58 \cdot 10^{-4}$
$q = 2^8$	$1.64 \cdot 10^{-8}$	$5.38 \cdot 10^{-8}$	$4.12 \cdot 10^{-8}$	$8.20 \cdot 10^{-8}$
$\epsilon = 0.16$				
$q = 2$	$2.54 \cdot 10^{-3}$	$1.33 \cdot 10^{-5}$	$4.63 \cdot 10^{-3}$	$5.92 \cdot 10^{-7}$
$q = 2^8$	$4.86 \cdot 10^{-7}$	$1.53 \cdot 10^{-10}$	$7.69 \cdot 10^{-7}$	$1.44 \cdot 10^{-12}$

low PER. As it will be seen in Section 6.5, the error introduced by this approximation in the pmf of S is negligible. Therefore, (6.7) is used hereafter.

Now we proceed to obtain the probability of successful content delivery S. For this, let $\boldsymbol{C}_{r\times c}$ be a coding matrix of dimension $r \times c$ s.t. $r \in \mathbb{N}$ and $\{c \in \mathbb{Z}_+ \mid c \leq g\}$, whose elements are selected uniformly at random from GF(q). The probability that matrix $\boldsymbol{C}_{r\times c}$ is full rank, denoted as $F(r, c)$, is

$$F(r,c) = \begin{cases} 0 & \text{for } r < c, \\ \prod_{j=0}^{c-1} \left(1 - q^{j-r}\right) & \text{otherwise.} \end{cases} \tag{6.9}$$

Then we use (6.9) to obtain the CDF of $T|X_0(i)$ as

$$F_{S|X_0}(s_i \mid x; i) = \sum_{u=g-x}^{s_i} \binom{s_i}{u} (1-\epsilon)^u \ \epsilon^{s_i-u} \ F(u, g-x) \tag{6.10}$$

which allows us to calculate the marginal CDF of $S(i)$,

$$\begin{aligned} F_S(s_i; i) &= \sum_{x=g_i}^{g} p_{X_0}(x; i) \ F_{S|X_0}(s_i \mid x; i) \\ &= \sum_{u=g}^{g+s_i} (1-\epsilon)^{u-g_i} \ \epsilon^{g+s_i-u} \sum_{x=x_{\min}}^{g} \binom{s_i}{u-x} \binom{g-g_i}{x-g_i} F(u-x, g-x) \end{aligned} \tag{6.11}$$

where $x_{\min} = \max\{g_i, u - s_i\}$.

To obtain the distribution of S, we first define the number of coded transmissions towards the ith UE s_i as a function of the number of time slots allocated for the transmission of coded packets s, as

$$s_i = f(s, i, n, g) = s + g_i - \left\lceil \frac{g + s - (i-1)}{n} \right\rceil. \tag{6.12}$$

That is, s_i transmissions will be performed by the UEs in $\mathcal{N}_i$ until time index s.

Tsimbalo *et al.* [101] define the probability of successful content delivery to be the probability that each and every UE in the MC decodes the generation. Let $S_{\mathcal{N}}$ be the RV that defines the probability of content delivery as defined by Tsimbalo *et al.*. The exact CDF of $S_{\mathcal{N}}$ is defined as

$$F_{S_{\mathcal{N}}}(s; n) \equiv \Pr\left[\bigcap_{i=1}^{n} X_{s_i}(i) = g\right]. \tag{6.13}$$

But obtaining the exact $F_{S_{\mathcal{N}}}(s; n)$ is complicated. Instead, it is commonly assumed that, at each s, $X_{s_i}(i) \perp\!\!\!\perp X_{s_j}(j)$ for all $\{i, j \in \mathcal{N} \mid i \neq j\}$. Hence, the lower bound

$$F'_{S_{\mathcal{N}}}(s; n) \equiv \prod_{i=1}^{n} \Pr\left[X_{s_i}(i) = g\right] = \prod_{i=1}^{n} F_{X_{s_i}}(g; i) \tag{6.14}$$

is commonly used. In particular, Tsimbalo *et al.* found (6.14) to be a tight lower bound for $F_{S_{\mathcal{N}}}(s; n)$ under the systematic RLNC for a wide range of values of q and g. Preliminary results on the performance of our NCC protocol were obtained with the definition of the probability of successful content delivery described previously [61]. Nevertheless, we find this definition to be inconvenient and unfair. Specifically, it is inconvenient in the sense that it is not exact, hence adopting this previous definition introduces an approximation error. We provide the following example to highlight the lack of fairness of the latter definition.

Please recall our main goal is to find s^*, defined as the minimum number of coded packet transmissions s needed to achieve the desired reliability τ. If we set the cloud size to be $n = 2$ and substitute F_S with $F'_{S_{\mathcal{N}}}$ in (6.1), we have

$$s^* \equiv \min_s \left\{s \mid \Pr\left[X_{s_1}(1) = g\right] \Pr\left[X_{s_2}(2) = g\right] \geq \tau\right\}. \tag{6.15}$$

Therefore, on average, the probability of decoding the generation at each UE must be $\geq \sqrt{\tau}$ when $n = 2$. Now assume $n = 100$. If we follow the same approach as described above, on average, the probability of decoding the generation at each UE must be $\geq \sqrt[100]{\tau}$ when $n = 100$. Since $\tau < 1$, the required $S(i)$ to reach the desired reliability grows with n. In other words, individual UEs in a small MC will have a lower probability to decode the generation than those in a larger MC. To solve this fairness problem, we propose to adopt the following definition for RV S.

Proposition 6.1. *Let S be the RV that defines the minimum probability that a given UE decodes the generation among the UEs in a specific MC at each coded packet transmission. That is, the minimum value of $S(i)$ among all the $i \in \mathcal{N}$ at each coded packet transmission. The CDF of S can be easily calculated as*

$$F_S(s) \equiv \min_i F_S(s_i; i) = \min_i \Pr\left[X_{s_i}(i) = g\right] \qquad \forall s \in \mathbb{N}. \tag{6.16}$$

Yet another interpretation of the proposed definition of S is the probability that the worst UE in the MC decodes the generation at each coded packet transmission. Please observe the proposed definition ensures a higher $S(i)$ for the rest of the UEs that is independent of the cloud size n.

Therefore, we use $F_S(s; n)$ as defined in (6.16) to calculate s^* for a given τ as in (6.1). Building on this, s^* defines the minimum number of coded packet transmissions needed so that $S(i) \geq \tau$, for all $i \in \mathcal{N}$. Hence, (6.1) can be rewritten as follows.

$$s^* \equiv \min_s \left\{ s \mid \min_i F_S(s_i; i) \geq \tau \right\}. \tag{6.17}$$

Once s^* has been obtained, we can calculate the maximum achievable throughput per UE R^*, given in bits per second. For this, let d be the length of the cellular phase in subframes. That is, the number of subframes needed to deliver the generation from the eNB to the MC. Since we assume no errors occur in the cellular phase, d is a deterministic value that highly depends on the multiplexing used for unicast data transmission. Therefore, we distinguish two main cases: TDM and FDM. In TDM, d only depends on the generation size g. On the other hand, in FDM, d depends on the generation size g, the cloud size n, the cellular data rate R, and the maximum

throughput in the carrier B. The latter depends on the selected MCS and the carrier bandwidth. Building on this, the length of the cellular phase is given as follows

$$d = \begin{cases} g & \text{for TDM,} \\ \left\lceil \dfrac{g}{\min\left\{n, \left\lfloor \frac{B}{R} \right\rfloor\right\}} \right\rceil & \text{for FDM.} \end{cases} \tag{6.18}$$

Please observe that the length of the cellular phase would be an RV if wireless channel errors can occur. However, this latter case can be easily extended from (6.18) as these errors do not affect the operation of our NCC protocol.

To proceed with the calculation of the achievable throughput, let ρ be the ratio of WiFi to cellular data rate. Oftentimes in 4G, only one interface can be used simultaneously. Hence, cellular and MC phases must be performed one after the other. As a result of this, the achievable throughput per UE in 4G is given as

$$R^*_{4\mathrm{G}}(n) = \frac{\ell}{t_s} \frac{g}{d + \frac{1}{\rho}(g + s^*)} = \frac{R}{\frac{d}{g} + \frac{1}{\rho}\left(1 + \frac{s^*}{g}\right)}, \qquad \text{if } n \geq 2. \tag{6.19}$$

On the other hand, 5G will provide full integration between cellular and short-range interfaces, so the cellular and MC phases can be performed in parallel after the first eNB transmission. As a result, the achievable throughput per UE in 5G can be calculated as

$$R^*_{5\mathrm{G}}(n) = \frac{\ell}{t_s} \frac{g}{1 + \frac{1}{\rho}(g + s^*)} = \frac{R}{\frac{1}{g} + \frac{1}{\rho}\left(1 + \frac{s^*}{g}\right)}, \qquad \text{if } n \geq 2. \tag{6.20}$$

That is, the MC phase can begin immediately after the first source data packet is transmitted from the eNB toward the MC in 5G. As a consequence, the multiplexing method used for unicast sessions is irrelevant.

Finally, we calculate the average energy consumption per UE $\overline{E}_{\mathrm{ue}}$. For this, let

$$\mathbb{E}\left[S(i) \mid s^*\right] = \sum_{u=0}^{f(s^*,i)} u \, p_S(u;i) \tag{6.21}$$

be the expected number of subframes that the ith UE is in reception mode and in which coded packets are transmitted. Please observe $p_S(u;i)$ is the pmf of $S_u(i)$, which can

be easily obtained from its CDF calculated by (6.11). We also calculate the expected number of source packets received at each of the UEs as

$$\mathbb{E}[X_0] = \frac{1}{n}\sum_{i=1}^{n}\sum_{x=0}^{g} x\, p_{X_0}(x;i) \tag{6.22}$$

From there we calculate $\overline{E}_{\text{ue}}(n)$ as

$$\overline{E}_{\text{ue}}(n) = \frac{\ell}{n}\Bigg[g\,E_{\text{cel,rx}} + (g+s^*)\,E_{\text{wifi,tx}} + \left((n-1)\,g + \sum_{i=1}^{n}\mathbb{E}\left[S(i) \mid s^*\right]\right)E_{\text{wifi,rx}} + \Big[s + n\,(g - \mathbb{E}[X_0])\Big]E_{\text{e/d}}(q)\Bigg] \tag{6.23}$$

where $E_{\text{cel,rx}}$, $E_{\text{wifi,rx}}$, and $E_{\text{wifi,tx}}$ define the energy consumed per bit in the LTE-A transmission, and WiFi reception and transmission, respectively. $E_{\text{e/d}}(q)$ is the energy consumed per bit to encode or decode a packet for a given Galois-field size GF(q). Please observe a different amount of energy can be consumed during encoding than during coding duties. However, the work of Sørensen *et al.* found the difference between these two to be negligible, hence it is safe to assume the same amount of energy is consumed [91].

6.5 Results

In this section we first compare the results obtained by our model with those obtained by Monte Carlo simulations and study the behavior of RV S. Second, we present the optimal number of coded transmissions s^* as a function of the cloud size n and discuss the achievable throughput gains. Finally, we evaluate the energy savings that can be achieved with our NCC protocol.

Throughout this section we select the generation size to be $g = 64$. The latter is one of the values that provide the highest benefits in NC [81, 91]; hence it is commonly used in the literature. Other common alternative is $g = 32$. Please recall the subframe duration in LTE-A is $t_s = 1$ ms. A typical UDP data packet of length $\ell = 1470$ bytes is used. Up to one data packet is transmitted to each UE in the MC per subframe, which

Table 6.3: Parameter settings.

Parameter	Symbol	Settings
Cell bandwidth	–	20 MHz
Generation size	g	64 packets
Galois-field size	q	$\{2, 2^8\}$
Cloud size	n	$\{2, 3, \ldots, 100\}$ UEs
Desired reliability	τ	$1 - 10^{-3}$
Packet erasure rate (PER)	ϵ	$\{0.2, 0.4, 0.8, 0.16\}$
Subframe duration	t_s	1 ms
Packet length	ℓ	1470 bytes
Data rate at the LTE-A and WiFi links	R	11.76 Mbps
Energy consumption for LTE-A reception [57]	$E_{\text{cel,rx}}$	78.68 nJ/bit
Energy consumption for WiFi transmission [93]	$E_{\text{wifi,tx}}$	37.64 nJ/bit
Energy consumption for WiFi reception [93]	$E_{\text{wifi,rx}}$	37.64 nJ/bit
Energy consumption for encoding/decoding [91]	$E_{\text{e/d}}\left(2^8\right)$	3.5 nJ/bit

gives a cellular data rate of $R = 11.76$ Mbps. We assume this same data rate for the WiFi links. Energy consumption parameters were obtained from the LTE-A and WiFi energy consumption models provided by Lauridsen *et al.* [57] and by Sun *et al.* [93], respectively. We assume the same energy per bit is consumed during transmission and reception over WiFi. The energy consumption during encoding and decoding is obtained from the work of Sørensen *et al.* [91], where a Samsung Galaxy S5 was considered. Other parameter settings are listed in Table 6.3.

A C-based simulator was developed to assess the accuracy of the analytical model; it comprises the coding, transmission, and encoding stages. The number of simulation runs is set to ensure the relative margin of error for each point of the pmf of successful content delivery S is less than 0.5 percent at a 95 percent confidence interval. For example, the minimum number of simulation runs to ensure the described accuracy was found to be one million for each combination of parameters.

Just as in Chapter 3, the accuracy of our model is assessed by means of the Jensen-

Table 6.4: JSD between the pmfs of successful content delivery obtained by our model and by simulations.

	$n = 3$		$n = 100$	
	$g = 32$	$g = 64$	$g = 32$	$g = 64$
$\epsilon = 0.02$				
$q = 2$	$3.49 \cdot 10^{-5}$	$6.85 \cdot 10^{-5}$	$5.26 \cdot 10^{-4}$	$4.68 \cdot 10^{-4}$
$q = 2^8$	$2.26 \cdot 10^{-5}$	$1.87 \cdot 10^{-5}$	$7.73 \cdot 10^{-4}$	$6.30 \cdot 10^{-4}$
$\epsilon = 0.16$				
$q = 2$	$1.21 \cdot 10^{-3}$	$1.10 \cdot 10^{-3}$	$1.17 \cdot 10^{-3}$	$1.42 \cdot 10^{-4}$
$q = 2^8$	$1.12 \cdot 10^{-4}$	$2.20 \cdot 10^{-4}$	$1.68 \cdot 10^{-4}$	$1.54 \cdot 10^{-5}$

Shannon Divergence (JSD), which measures the increase in the Shannon's entropy when an approximated pmf is assumed to be the real pmf of an RV. The formula to calculate the JSD between two pmfs was defined in (3.53). To calculate the JSD, we denote the pmfs of S obtained by our model and by simulation as $p_S(s;n)$ and $p_{S_{\text{sim}}}(s;n)$, respectively, to explicitly indicate these are calculated for a specific value of n. Building o this, (3.53) can be written as

$$\text{JSD}\left(p_S(s;n)\right) \equiv H\left(\frac{p_{S_{\text{sim}}}(s;n) + p_S(s;n)}{2}\right) - \frac{H\left(p_{S_{\text{sim}}}(s;n)\right) + H\left(p_S(s;n)\right)}{2}$$

where $H(\cdot)$ is the base-e Shannon's entropy. As such, the JSD is upper bounded by $\log 2$ and a JSD of zero indicates both pmfs are identical. Hence, $0 \leq \text{JSD}(\cdot) \leq \log 2$.

Table 6.4 shows the JSD between $p_S(s;n)$ and $p_{S_{\text{sim}}}(s;n)$ for typical values of $g \in \{32, 64\}$, $q \in \{2, 2^8\}$, and for widely distinct values of $n \in \{3, 100\}$ and $\epsilon \in \{0.02, 0.16\}$. As it can be seen, the JSD is extremely low regardless of the cloud and generation sizes. As a reference, the obtained JSD for $n = 2$, where our formulations are exact, in combination with $g = 64$, $q = 2^8$, and $\epsilon = 0.02$ is $2.145 \cdot 10^{-4}$. The JSD for the same combination of n, g, and q, but with $\epsilon = 0.16$ is $2.65 \cdot 10^{-3}$. One million simulations were performed for these cases, hence, we consider all cases that lead to a comparable of lower JSD to be exact.

We begin the analysis of our NCC protocol by comparing the complementary

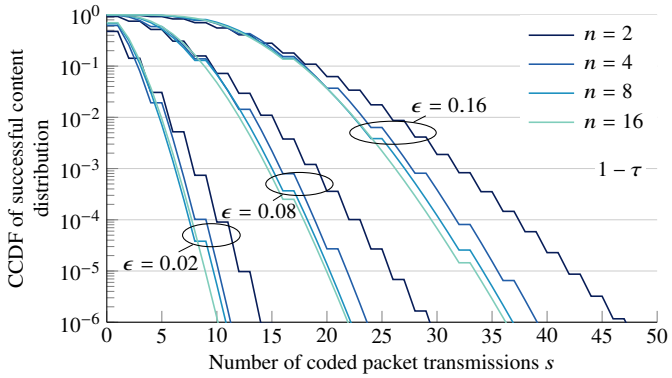


Figure 6.6: CCDF of successful content delivery S for $q = 2^8$, $\epsilon = \{0.02, 0.08, 0.16\}$, and $n = \{2, 4, 8, 16\}$; y-axis in logarithmic scale.

CDF (CCDF) of successful content delivery $1 - F_S(s; n)$, for $n = \{2, 4, 8, 16\}$ and $\epsilon = \{0.02, 0.08, 0.16\}$ in Fig. 6.6. In other words, Fig 6.6 shows the probability that the worst UE in the MC does not decode the generation. Therefore, lower values indicate a better performance. As it can be seen, large cloud sizes usually reduce the number of coded packet transmissions needed to achieve the desired reliability. The reason for this is that the ratio of transmissions from the UEs in $\mathcal{N}_i$ to total transmissions in the MC phase s_i/s^* increases with n. In other words, the frequency of the packets transmitted in the MC towards each of the n UEs increases with n. This can be clearly seen in the plots for $n = 2$. In these, the UEs receive up to one packet every two subframes and transmit in the other. Hence, these can receive up to 50 percent of the transmitted packets. This effect can be observed in the step-like shape of these curves. Instead, UEs in an MC for which $n = 16$ receive 15 data packets every 16 subframes. Hence, these receive up to 93.75 percent of the transmitted packets. In the following, we obtain s^* for $\tau = 1 - 10^{-3}$.

The effect of cloud size on performance can be clearly observed in Fig 6.7, where we show s^* and the achieved throughput per UE as a function of n and ϵ for $q = 2^8$. The results presented in Fig. 6.7b were obtained under the assumption that TDM is used for the unicast sessions under 4G. That is, data packets are transmitted one after

the other and the MC phase begins at the end of the cellular phase. Specifically, the selection of $n = 2$ results in the largest s^* and the lowest throughput. Clearly, s^* is not a monotonically decreasing function. Nevertheless, it is clear that increasing the cloud size will oftentimes be beneficial. For instance, selecting $n \geq 7$ and $n \geq 10$ is optimal for $\epsilon = 0.02$ and for $\epsilon = 0.04$, respectively. On the other hand, selecting $15 \leq n \leq 32$ and $26 \leq n \leq 32$ is optimal for $\epsilon = 0.08$ and for $\epsilon = 0.16$, respectively. Interestingly, the global minima of s^* for each of the values of $\epsilon \in \{0.02, 0.04, 0.08, 0.16\}$ are $\{7, 10, 15, 26\}$, which coincides with the values of n that lead to these values. Also of interest is that, if $q = 2$ were to be selected instead of $q = 2^8$, the achievable throughput would be reduced around 6 and 3 percent for $\epsilon = 0.02$ and for $\epsilon = 0.16$, respectively.

Please observe that the maximum achievable throughput per UE within the MC using TDM in 4G is lower than that of a single unicast session $R = 11.76$ Mbps. For example, $R_{\text{ue}}(n) \approx R/2.15$ for all n given $\epsilon = 0.08$. This slight decrease in throughput can be seen as the main overhead of our NCC protocol, and, as described by (6.19), occurs because g packet transmissions are performed in the cellular phase, followed by g systematic and s^* coded transmissions in the MC phase.

Nevertheless, this decrease in throughput can only occur if the cellular bandwidth is sufficient to allocate n unicast sessions in parallel, which is needed to maintain the throughput per UE equal to the data rate in traditional unicast content delivery. On the other hand, our NCC protocol can provide throughput gains when the cellular bandwidth is not sufficient. For this, let the cellular bandwidth be 20 MHz, which is the maximum bandwidth of an LTE-A carrier. If the highest MCS of 256 quadrature amplitude modulation (QAM) is used, the maximum data rate that can be achieved in such carrier is $B = 97.896$ Mbps [8]. Therefore, the maximum throughput per UE following the traditional approach is $R_{\max}(n) = \min\{R, 97.896/n\}$. Building on this, Fig. 6.8 shows the achievable throughput gains per UE with our NCC protocol for a given n, defined as

$$G_{\text{th}}(n) = \frac{R^*(n)}{R_{\max}(n)} - 1. \tag{6.24}$$

Three cases are considered in Fig. 6.8 with respect to cellular data transmission: 1) TDM in 4G; 2) FDM in 4G; and 3) 5G. Please recall that we assume only one interface at the time can be used in 4G, so cellular and MC phases must be performed in

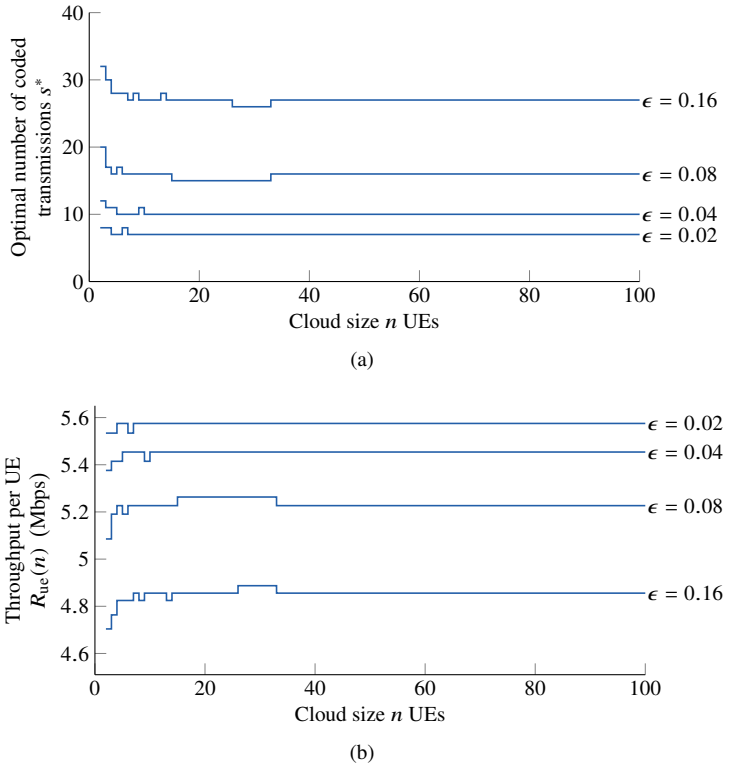


Figure 6.7: (a) Optimal number of coded packet transmissions s^* and (b) throughput per UE given TDM is used for the unicast sessions under 4G; $\tau = 1 - 10^{-3}$ and $q = 2^8$.

sequence. On the other hand, these phases can be performed in parallel in 5G, hence, the multiplexing method is irrelevant to the calculation of the throughput.

As it can be seen, throughput losses occur with small cloud sizes and the reason for this was described above. However, throughput gains are obtained if $n \geq 21$ and if $n \geq 14$ for the TDM and FDM methods in 4G. On the other hand, throughput gains are obtained if $n \geq 12$ in 5G. It is important to mention that these results were obtained with $\epsilon = 0.16$, which represents a high PER and can be even be seen as an upper bound for this parameter. In other words, these results may be seen as a worst case scenario for WiFi transmissions.

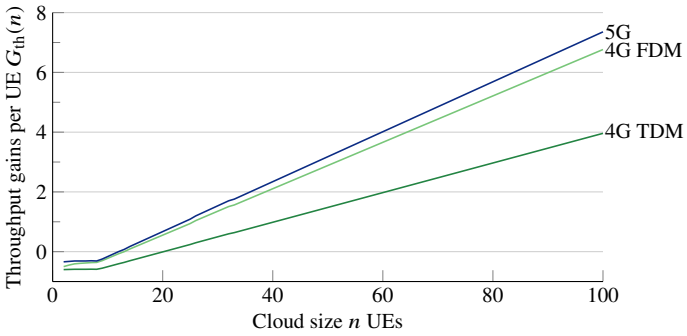


Figure 6.8: Achievable throughput gains with our NCC protocol given: 1) TDM in 4G; 2) FDM in 4G; and 3) 5G; $\epsilon = 0.16$.

Clearly, an inherent benefit of our NCC protocol is offloading the cellular link. This in turn results in the added benefit of cellular data savings for the UEs in the MC. Naturally, UEs in an MC only download a fraction of the data from the cellular link, which is inversely proportional to n. Hence, the cellular data savings for the UEs can be easily calculated as $1 - 1/n$.

Finally, we present the sharp reduction in the energy consumption (i.e., energy savings) that can be achieved at the UEs. For this, Fig 6.9 shows an area plot of the average energy consumption per UE as a function of n for $\epsilon = 0.16$. Colors indicate the energy consumption at each interface and process, namely LTE-A reception, WiFi reception and transmission, and encoding/decoding. For example, the energy consumption for the direct transmission of the g packets to each UE through the cellular link is 59.17 mJ. On the other hand, the energy consumption per UE for $n = 20$ is 37.47 mJ and is further reduced as n increases. Therefore, energy savings of more than 37 percent can be achieved with our NCC protocol, even with relatively small cloud sizes and a high PER.

Fig. 6.9 also shows that the main contributing factor to the overall energy savings is that the number of packets transmitted from the eNB to each UE decreases as n increases. Conversely, the number of packets transmitted through WiFi to each UE increases with n, but the power consumption during reception in the WiFi link is

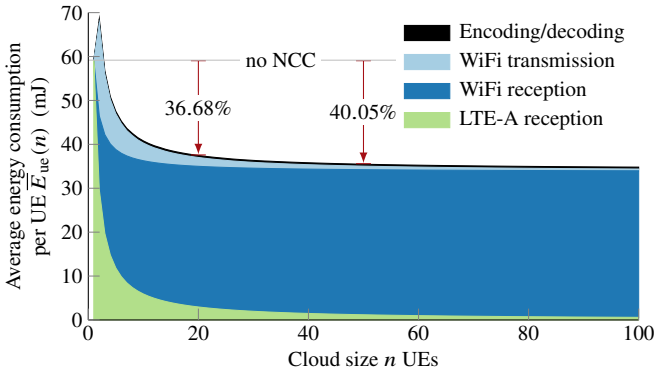


Figure 6.9: Average energy consumption per UE given $\epsilon = 0.16$ and $q = 2^8$.

much lower than in the cellular link. The energy consumed for WiFi transmissions becomes particularly small for large cloud sizes. Finally, the energy consumption during encoding and decoding is, in general, the least contributing factor to overall energy consumption despite it slightly increases with n. Nevertheless, the latter surpasses the average energy consumed during WiFi transmissions for $n \geq 94$.

6.6 Conclusions

In this chapter, we presented an NCC protocol for massive content delivery and a simple but accurate analytical model that allows to fine-tune its parameters. Specifically, the presented analytical model was used to find the optimal configuration for the NCC protocol. Our analytical model incorporates an upper bound for the probability of linear independence of coded packets that sharply reduces its complexity. We evaluated the error of this upper bound and observed that its impact is negligible when typical generation and Galois-field sizes are selected.

Our results show that important energy savings of more than 37 percent can be achieved with our protocol, even with relatively small cloud sizes and considerably high PERs. The main overhead of our protocol is the decrease in throughput when compared to data transmission through parallel unicast LTE-A links. But an eNB can

only serve a limited number of unicast sessions in parallel at a high data rate (or at any data rate if the number of UEs is extremely large). Hence, important throughput gains are achieved when the cellular bandwidth is insufficient to serve the requesting UEs at the desired data rate. In our studies, this occurs when more than 14 UEs request the same content and are served with a single LTE-A (i.e., 4G) carrier. In 5G, throughput gains can be achieved with only 12 UEs or more. In addition to energy savings and throughput gains, our NCC provides dramatic cellular data savings that grow with the number of UEs in the MC. The combination of these benefits makes our NCC protocol an appealing solution under massive content delivery scenarios.

A relevant characteristic that is not captured by our model is that the PER between some pairs of UEs increases with the cloud size in practical implementations. As a result, important differences in the PER between pairs of UEs are expected if large MCs are formed. Hence, the achievable throughput will be limited by the maximum PER in the MC. Building on this, we advise to set the minimum cloud size that results in the maximum throughput. By doing so, small MCs will be formed with closely-located UEs and, still, important energy savings will be achieved.

Chapter 7

Conclusions and future perspectives

This thesis mainly focuses on efficiently supporting massive machine-type communication (mMTC) in cellular networks, but also incorporates analyses on mMTC over wireless sensor networks (WSNs) and a potential solution to achieve efficient enhanced mobile broadband (eMBB) in cellular networks under massive content delivery scenarios. As such, it presents analyses and solutions to two out of the three main use cases for the 5th generation (5G) of mobile networks. 5G networks promise full integration with 4th generation (4G) and short-range technologies, and will adopt several aspects of 4G directly. Therefore, the analyses presented in this thesis that assume the typical configuration in 4G can be immediately extended to 5G.

The results presented in Chapters 2, 3, and 4 exhibited the inefficiency of the RA procedure (RAP), defined by the 3rd Generation Partnership Project (3GPP) for the initial access to cellular networks, under mMTC. The RAP is a four-message handshake between the user equipments (UEs) and the cellular base station, whose first step: preamble transmission, suffers from the same limitations as a simple multichannel slotted ALOHA protocol. In addition, downlink signaling resources used to signal the success of preamble transmissions present a second limitation that can be even more restrictive. As a result of this, the probability that a UE successfully completes the RAP (i.e., success probability) in LTE Advanced (LTE-A) under the most typical mMTC behavior is only 0.313. This situation becomes even more critical for the

narrowband Internet of Things (NB-IoT) standard, simply because the number of available preambles is lesser when compared to traditional LTE-A. For example, it was observed in Chapter 4 that the success probability under an mMTC scenario with 30 available preambles is barely 0.115. These values are far from the target success probability, defined to be 0.95.

A C-based simulator and an analytical model of the RAP were developed to evaluate the performance of the RAP and to investigate possible solutions to the problem of congestion under mMTC scenarios. The analytical model described in Chapter 3 is among the most accurate models that can be found in the literature. For instance, the maximum relative error obtained with our model with respect to simulations in the success probability was 0.36 percent. Furthermore, our model can be easily adapted to different assumptions with respect to the handling of preamble collisions; this characteristic was showcased in Section 3.3.5. This analytical model is one of the main contributions of this thesis.

Simulation and analytical results, presented in Chapter 2 and Chapter 3, respectively, revealed that severe congestion occurs whenever the number of accesses per random access opportunity (RAO) exceeds the capacity of the RAP; this capacity is calculated by (2.8). Initial efforts were focused on increasing the success probability by fine-tuning the configuration parameters of the system. These include: increasing the frequency of RAOs, increasing the number of available preambles, reducing the maximum number of access attempts per UE, and replacing the uniform backoff defined in the standards with an exponential backoff. Our results were conclusive: even though the manipulation these parameters can lead to a higher success probability, the target success probability of 0.95 can only be obtained with an exceedingly large number of available preambles. This is unattainable in the practice and implies that the target success probability can only be obtained by implementing an access control scheme.

The access class barring (ACB) scheme is an access control scheme defined in the 3GPP standards that uses a probabilistic approach to redistribute the access attempts through time. The benefits of the ACB scheme with fixed parameters were investigated by simulation in Chapter 2 and by the analytical model in Chapter 3. Our results

show that the parameters of the ACB scheme can be optimized according to the signaling traffic characteristics. That is, these can be tailored to achieve a success probability higher than 0.95 with the minimum access delay. However, the minimum 95th percentile of access delay that can be achieved with an ACB scheme with fixed parameters is 13.58 s with the uniform backoff and 11.24 s with the exponential backoff. These values are exceedingly long, even for delay-tolerant applications. Nevertheless, our results suggested that access delay can be considerably reduced by adapting the ACB parameters to the signaling traffic intensity in real time.

Chapter 4 presented an efficient and practical solution to congestion under mMTC scenarios in the form of an adaptive mechanism to automatically adapt the ACB parameters to the intensity of accesses: an access class barring configuration (ACBC) scheme. This ACBC scheme can efficiently relieve congestion under mMTC scenarios and is one of the few solutions reported in the literature that adheres to the 3GPP standards. Therefore, it can be directly implemented at the cellular base stations. This scheme relies on the ratio of idle to available resources as the main load indicator, whose sudden variations are suppressed by an adaptive filter. The result is a much more stable output when compared to the case with no adaptive filter. Our results show that the ACBC scheme can reduce the 95th percentile of access delay by up to 50 percent when compared to the optimal configuration of the ACB with fixed parameters. Yet another relevant characteristic of the ACBC scheme is that it is able to maintain a near-optimal performance even when one of its main configuration parameters, the barring indicator, is selected to be higher than the optimal value. Specifically, this inaccurate configuration only results in a slight increase in the access delay.

Future work in the area of mMTC in cellular networks includes the update of the simulator and of the analytical model to new enhancements to the RAP that may be introduced in the second standardization phase of 5G. This would provide a reliable picture to the impact of these new enhancements on the capacity of the new RAP. Yet another interesting line of research is that of cooperative random access (RA) approaches. For instance, the formulation of grouping or clustering algorithms that are specific to cellular networks can efficiently prevent congestion and provide a shorter access delay than access control mechanisms. For example, for the scenario studied in Chapters 2 to 4, the creation of groups with only three UEs can reduce the signaling

traffic intensity below the capacity of the RAP and prevent congestion.

On the other hand, while the ACBC scheme presented in Chapter 4 leads to remarkably positive results, it may give the impression that too many configuration parameters are involved. The fact that some of these parameters must be selected empirically is one of the main contributing factor to this perception. While all the configuration parameters involved in our ACBC scheme were studied in this thesis and their adequate or optimal values were presented, slightly different parameters may be needed when the target mMTC scenario is widely different to the traffic model (TM) 2. Sadly, the need to select some configuration parameters empirically is indeed one of the main inherent drawbacks of adaptive filters, so little to no refinements can be made to our scheme by following this same approach.

It is important to recall that, although other adaptive algorithms such as the recursive least-squares (RLS), along with fixed filters, were studied, these were outperformed by the least-mean-square (LMS). This is the reason the LMS is implemented in our ACBC scheme. On the other hand, recent advances in machine learning techniques can be incorporated to our ACBC scheme. That is, to use the same mechanism to calculate the load indicator, described in Chapter 4 (i.e., the ratio of idle to available resources), as an input to a machine learning algorithm that selects the highest possible barring rate to achieve the desired success probability. Such approach may yield to similar results as the ones presented in this thesis, but with fewer configuration parameters that determine the efficacy of the scheme.

Chapter 5 provided a different perspective to mMTC. Namely, Chapter 5 investigated RA protocols for event reporting in cluster-based WSNs. This is a more general and traditional approach to machine-type communications (MTC) than the one covered in Chapters 2 to 4. Specifically, a hybrid model that combines simulation results with analytical modeling was presented. The results provided by this model exhibited two approaches that greatly enhance the performance of RA event reporting. The first one is to set a maximum of event reports to be transmitted per cluster and instruct nodes to overhear packet transmissions. By doing so, redundant packets can be discarded. The second one is to reduce transmission probabilities when a collision occurs. The latter approach, known as the adaptive backoff (AB), is certainly intuitive and has been

explored in the literature, but its combination with the first approach intensifies its benefits. For instance, our results show that a similar energy consumption and access delay can be achieved with and without the AB, but network performance is much more robust with the AB. That is, the network performance is much more resilient to the inaccurate selection of transmission probabilities. For instance, a relative error of more than 40 percent can be made in the selection of the transmission probabilities and still achieve the minimum 90th percentile of access delay. Whereas such an error without the AB causes a one-fold increase in the 90th percentile of access delay. It is important to emphasize that the guidelines provided in Chapter 4 can be easily applied to more complex WSN protocols, but also in other related systems of similar nature. For example, these guidelines can be applied to the formation of clusters or groups for the cooperative RA to cellular networks. In such an approach, the few first UEs that successfully complete the RAP under an mMTC application can inform their status to neighboring UEs and serve as cluster heads (CHs). Then, neighboring UEs would compete to join the cluster with the closest CH.

Finally, Chapter 6 complements the analyses on mMTC by providing an energy efficient solution to eMBB in cellular networks. The proposed solution is based on network-coded cooperation (NCC), which is the combination of network coding (NC) with cooperative architectures known as mobile clouds (MCs). Specifically, Chapter 6 presented an NCC protocol to offload the cellular link under massive content delivery scenarios. In this protocol, only a fraction of the data is sent to each of the UEs in the cloud. Then, these UEs cooperate through multicast WiFi links to distribute the data among the MC. In addition, an analytical model of this protocol was formulated; with this model, the minimum number of coded packet transmissions to achieve a target reliability was calculated. Our results show that the potential gains offered by this NCC protocol when compared to traditional content replication through independent unicast cellular links are numerous. For instance: 1) energy savings can exceed 37 percent, even with relatively small cloud sizes of 20 UEs; 2) throughput gains that increase linearly with the number of UEs can be achieved when the cellular bandwidth is insufficient; and 3) cellular data savings are proportional to the number of UEs in the MC.

Future work related to NCC includes the refinement of the analytical model and

the real-life implementation of the NCC protocol in current smart phones, but also the extension of this work to other scenarios in which the benefits of diverse wireless interfaces are exploited. This topic has acquired great significance as one of the priorities of 5G is to provide a deep level of integration between cellular and short-range interfaces and its potential benefits are numerous. Some of the possible applications are: coverage extension in rural areas, cooperative massive access, platooning, vehicular networks, and many more.

Appendices

Appendix A

Notations

X	Upper case symbols represent random variables
X_s	A lower case subindex in a random variable represents the time index
X_S	An upper case subindex in a random variable serves as an additional identifier
$\{X_s\}$	A random variable with time index and between braces represents a stochastic process
$\boldsymbol{x}$	Boldface lower case symbols represent vectors
$\boldsymbol{X}$	Boldface upper case symbols represent matrices
$\mathcal{X}$	Calligraphic symbols represent sets and events
X	Upper case roman symbols represent frequently used functions
x^*	An asterisk in the superscript represents the optimal value of a given variable
$p_X(x)$	Probability mass function (pmf) of random variable X whose domain is x
$F_X(x)$	Cumulative distribution function (CDF) of random variable X whose domain is x
P_a	Probability that event a occurs; only used when frequently repeated
$\Pr[X = x]$	Probability that random variable X is equal to x
$\mathbb{R}$	Set of real numbers
$\mathbb{N}$	Set of natural numbers, including 0
$\mathbb{Z}_+$	Set of positive integers

Appendix B

Derivations

B.1 Lower bounds for the physical RACH (PRACH) capacity

Simple lower bounds for the PRACH capacity $\mathrm{C}(r)$ are derived in this Appendix. These were introduced in (2.6) and (2.7). For this, please recall that $n^*(i) = \left[\log\left(r/\left[r-1\right]\right)\right]^{-1}$ is the number of contending UEs that maximizes the expected number of successful preambles $\mathbb{E}[S]$ at the ith RAO. Next, from inequalities

$$1 - \frac{1}{a} < \log(a) < a - 1 \qquad \text{for } a > 0 \tag{B.1}$$

we obtain

$$\left(\frac{r}{r-1} - 1\right)^{-1} < \left[\log\left(\frac{r}{r-1}\right)\right]^{-1} < \left(1 - \frac{r-1}{r}\right)^{-1} \tag{B.2}$$

which gives

$$r - 1 < n^*(i) < r. \tag{B.3}$$

By applying the inequalities in (B.3) to (2.3), which defines the PRACH capacity and is

$$\mathrm{C}(r) = \max_{n(i)} \mathbb{E}[S] = \left[\log\left(\frac{r}{r-1}\right)\right]^{-1}\left(1 - \frac{1}{r}\right)^{\left[\log\left(\frac{r}{r-1}\right)\right]^{-1}-1}$$

we obtain

$$r\left(1-\frac{1}{r}\right)^{r} < \mathrm{C}\,(r) < r\left(1-\frac{1}{r}\right)^{r-2}. \tag{B.4}$$

From there, it can be easily seen that

$$r\left(1-\frac{1}{r}\right)^{r} < r\left(1-\frac{1}{r}\right)^{r-1} < r\left(1-\frac{1}{r}\right)^{r-2}, \qquad \text{for } r>0; \tag{B.5}$$

therefore, $r\,(1-1/r)^{r-1} \approx \mathrm{C}\,(r)$.

Now, by observing that both, the increasing function $(1-1/r)^{r} \to e^{-1}$ and the decreasing function $(1-1/r)^{r-1} \to e^{-1}$ as $r \to \infty$ we can see that

$$r\left(1-\frac{1}{r}\right)^{r} < \frac{r}{e} < r\left(1-\frac{1}{r}\right)^{r-1}. \tag{B.6}$$

Therefore, $r\,(1-1/r)^{r-1}$ and r/e are lower bounds for $\mathrm{C}\,(r)$ that correspond to (2.6) and (2.7), respectively.

B.2 Proof of Lemma 3.1.

This appendix presents the proof of Lemma 3.1.

Proof. As described by (3.1) on page 47, the continuous-time Beta distribution is defined by numerator $t^{\alpha-1}\,(1-t)^{\beta-1}$ and the denominator $\mathrm{B}\,(\alpha,\beta)$ merely serves as a normalization constant. Building on this, the probability mass function (pmf) of a discrete time random variable (RV) $T_d \sim \mathrm{Beta}\,(\alpha,\beta)$ can be defined as follows.

$$\begin{aligned} p_{T_d}\left(\frac{i}{i_{\mathrm{dist}}};\alpha,\beta\right) &= \frac{\left(\frac{i}{i_{\mathrm{dist}}}\right)^{\alpha-1}\left(1-\frac{i}{i_{\mathrm{dist}}}\right)^{\beta-1}}{\sum_{i=0}^{i_{\mathrm{dist}}}\left(\frac{i}{i_{\mathrm{dist}}}\right)^{\alpha-1}\left(1-\frac{i}{i_{\mathrm{dist}}}\right)^{\beta-1}} \\ &= \frac{i^{\alpha-1}\,(i_{\mathrm{dist}}-i)^{\beta-1}}{\sum_{i=0}^{i_{\mathrm{dist}}} i^{\alpha-1}\,(i_{\mathrm{dist}}-i)^{\beta-1}} \end{aligned} \tag{B.7}$$

For $\alpha = 3$ and $\beta = 4$ as defined for the TM 2, the numerator in (B.7) becomes

$$\sum_{i=0}^{i_{\text{dist}}} i^2 \left(i_{\text{dist}} - i\right)^3 = \frac{i_{\text{dist}}^6 - i_{\text{dist}}^2}{60} = \frac{i_{\text{dist}}^6 - i_{\text{dist}}^2}{\mathrm{B}\left(3,4\right)} \tag{B.8}$$

which gives

$$p_{T_d}\left(\frac{i}{i_{\text{dist}}}; 3, 4\right) = \frac{60 i^2 \left(i_{\text{dist}} - i\right)^3}{i_{\text{dist}}^6 - i_{\text{dist}}^2}. \tag{B.9}$$

□

Appendix C

Performance of the proposed ACBC scheme with the RLS algorithm

The process to configure our ACBC scheme, presented in Chapter 4, with the RLS algorithm and its potential benefits are presented in this Appendix. As with the LMS algorithm, the RLS is to be implemented at the "Adaptive filtering" block depicted in Fig. 4.1 on page 92.

The RLS belongs to a different family of adaptive algorithms to that of the LMS. That is, the RLS is a recursive implementation of the method of *least squares*, whereas the LMS is an application of the method of *stochastic gradient descent*. Both of these algorithms have advantages and disadvantages. In particular, the rate of convergence of the RLS is up to an order of magnitude faster than that of the LMS [45]. On the other hand, the LMS is less sensitive to disturbances (i.e., more robust) and less complex (computationally speaking) than the RLS. In particular, the complexity of the RLS algorithm is on the order of $O\left(\ell^2\right)$, which is higher than that of the LMS algorithm: $O\left(\ell\right)$. Nevertheless, the mathematical formulation and implementation of the RLS is relatively simple [45].

Like the LMS, the RLS adaptive filter algorithm, summarized in Algorithm 6, consists of a filtering and an adaptive process.

It is essential to observe the difference between the *a priori* estimation error $\xi(j)$,

Algorithm 3 RLS adaptive algorithm.

Require: the number of filter coefficients ℓ

Require: regularization parameter $\delta > 0$

Require: forgetting factor λ

1: Initialize the vector of filter coefficients $\boldsymbol{w}\,(0)$ and the input vector $\boldsymbol{u}\,(0)$ as

$$w_m\,(0) = u\,(-m) = 0, \qquad m \in \{0, 1, \ldots, \ell - 1\} \tag{C.1}$$

2: Initialize the inverse correlation matrix $\boldsymbol{P}\,(0) = \delta^{-1}\,\boldsymbol{I}$

3: **for all** $j = 1, 2, \ldots$ **do**

4: Filtering process:

$$y(j) = \boldsymbol{w}^{\mathsf{T}}(j-1)\boldsymbol{u}(j) \tag{C.2}$$

5: Adaptive process:

$$\boldsymbol{k}(j) = \frac{\lambda^{-1}\,\boldsymbol{P}\,(j-1)\;\boldsymbol{u}(j)}{1 + \lambda^{-1}\,\boldsymbol{u}^{\mathsf{T}}(j)\,\boldsymbol{P}\,(j-1)\;\boldsymbol{u}(j)} \tag{C.3a}$$

$$\xi(j) = d(j) - \boldsymbol{w}\,(j-1)\;\boldsymbol{u}(j) \tag{C.3b}$$

$$\boldsymbol{w}\,(j) = \boldsymbol{w}(j-1) + \boldsymbol{k}(j)\xi(j) \tag{C.3c}$$

$$\boldsymbol{P}\,(j) = \lambda^{-1}\,\boldsymbol{P}\,(j-1) - \lambda^{-1}\,\boldsymbol{k}(j)\,\boldsymbol{u}^{\mathsf{T}}(j)\,\boldsymbol{P}\,(j-1) \tag{C.3d}$$

6: **end for**

calculated in the RLS algorithm and the *a posteriori* estimation error $e(j)$, calculated in the LMS algorithm. The latter is the difference between the desired response $d(j)$ and the output of the filter at time j. Conversely, the former represents an estimate of the desired response $d(j)$ based on the old least-squares estimate of $\boldsymbol{w}\,(j-1)$. Please refer to [45, Chapter 10] for a thorough discussion on this matter.

The selected configuration of the RLS algorithm for our ACBC scheme is analogous to that of the "pulling" ALE (PALE) configuration with the LMS algorithm depicted in Fig. 4.2, hence, the desired response is set to be $d(j) = 1$. Fig. C.1 illustrates the

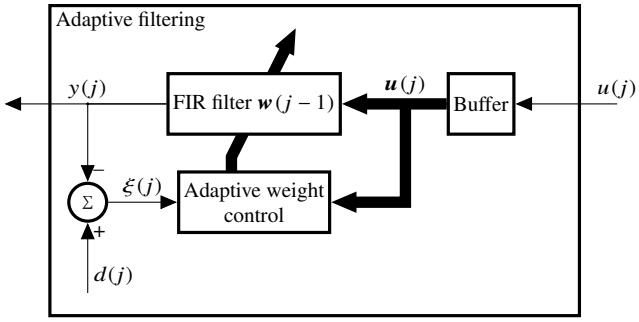


Figure C.1: Block diagram of the RLS adaptive filter algorithm.

internal structure of the algorithm.

The RLS algorithm has two parameters that must be selected empirically, namely the regularization parameter δ (please observe this is different to function $\delta(i)$) and the forgetting factor λ (this parameter is also different to the one used in Chapter 2, which refers to the access intensity of human-to-human (H2H) UEs). The latter determines the rate at which the algorithm "forgets" the previous inputs. As with parameter μ for the LMS, there is no exact method to select δ nor λ. However, the following recommendations exist.

- Regularization parameter δ: Select a small positive value when the input (in our case $u(j)$) is relatively high with respect to its sudden variations. Select a large positive value otherwise.
- Forgetting factor λ: Select a positive value that is close to, but less than, 1. This ensures past values of the input $u(j)$ are forgotten by the algorithm. On the other hand, $\lambda = 1$ corresponds to the method of least squares.

Therefore, an adequate value for these parameters must be selected by observing the response of the algorithm. We find adequate values for these parameters in an analogous process as the one described in Chapter 4. That is, we disable the ACB scheme and aim to identify the values that successfully suppress the sudden variations of $u(j)$, but with the fastest possible convergence of the barring rate $p_{\text{acb}}(j)$ to $\mathbb{E}\left[u(j)\right]$.

In Fig. 4.5 we observed that the variations of $u(j)$ are relatively low when compared to $\mathbb{E}[u(j)]$ under the the TM 1. Building on this, we know we need to select a small positive value for δ.

To find adequate values for δ and λ, we observed the response of the algorithm with $\delta \in \{0.001, 0.01, 0.1, 1\}$ and $\lambda \in \{0.99, 0.999, 0.9999\}$. Fig. C.2 illustrates the response of the algorithm with all possible combinations of these values given $r = 54$, along with the calculated $u(j)$. Clearly, the response is not adequate for $\lambda = 0.99$. Also, the difference in the response between selecting $\lambda = 0.999$ and $\lambda = 0.999$ is negligible, and a similar behavior was observed for higher values of λ.

Fig. C.2 also shows that the response is highly variable with $\delta = 1$ and the slowest convergence occurs with $\delta = 0.001$; the latter value also causes the response to be lower than the expected value of $u(j)$ under the traffic model 1 $\mathbb{E}[u(j)] \approx 5/6$. Please refer Chapter 4.6 for details on the calculation of this value. On the other hand, it is difficult to determine whether the best response is provided by selecting $\delta = 0.1$ or $\delta = 0.01$. Therefore, we select the latter value for the following tests.

Once we have identified the adequate values for $\lambda = 0.999$ and $\delta = 0.01$, we can find the optimal performance of our ACBC scheme with theRLS algorithm. Please observe that the calculated $p_{\text{acb}}(j)$, as shown in Fig. C.2, is always lower than 1, so we must define $\omega > 0$ to avoid an unnecessary delay under the TM 1. In particular, we select $\omega = 3$.

With these values, we found the optimal configuration to be $\ell^* = 32$ and $t_{\max} = 1$ s, which led to $P_s = 0.962$, $D_{95} = 7.531$ s, and $\mathbb{E}[K] = 2.415$, given $r = 54$. The values reported in Chapter 4.6 with the adaptive line enhancer (ALE) and PALE configurations are $D_{95} = 6.807$ s and $D_{95} = 7.286$ s, respectively.

Finally, Fig. C.3 shows P_s and D_{95} under the TM 2 for our ACBC scheme with the RLS algorithm given $\ell \in \{4, 8, 16, 32\}$ and $t_{\max} \in \{0.1, 0.2, \ldots, 5\}$ s. The minimum D_{95} obtained with the ALE configuration, denoted as ALE*, is also illustrated. From this Fig. C.3 it can be seen that the behavior of our ACBC with the RLS is comparable to that of the LMS, but it is less robust. For example, the P_s obtained with $\ell = 16$ increases to ≈ 1 by $t_{\max} = 1$, but then drops slightly for $1 < t_{\max} < 4$. These results confirm the superior benefits are provided by the LMS algorithm.

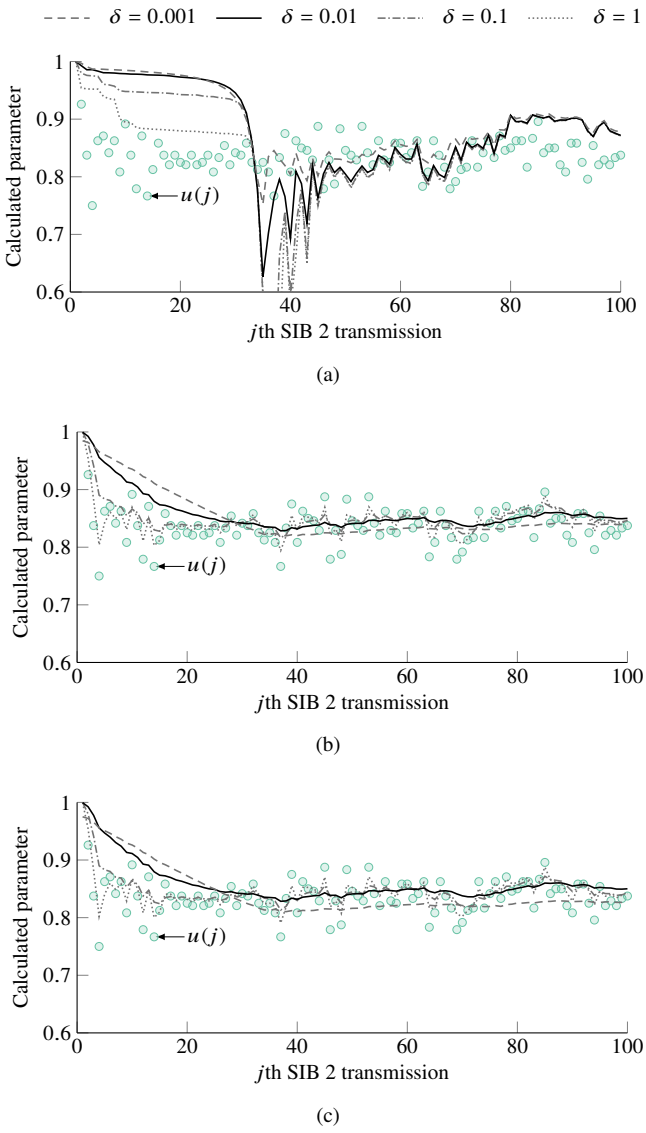


Figure C.2: Ratio of idle to available resources $u(j)$ and barring rate $p_{\text{acb}}(j)$ calculated at the jth SIB2 for a single simulation run and $r = 54$ for the RLS algorithm with $\delta \in \{1, 0.1, 0.01, 0.001\}$ and (a) $\lambda = 0.99$, (b) $\lambda = 0.999$, and (c) $\lambda = 0.9999$.

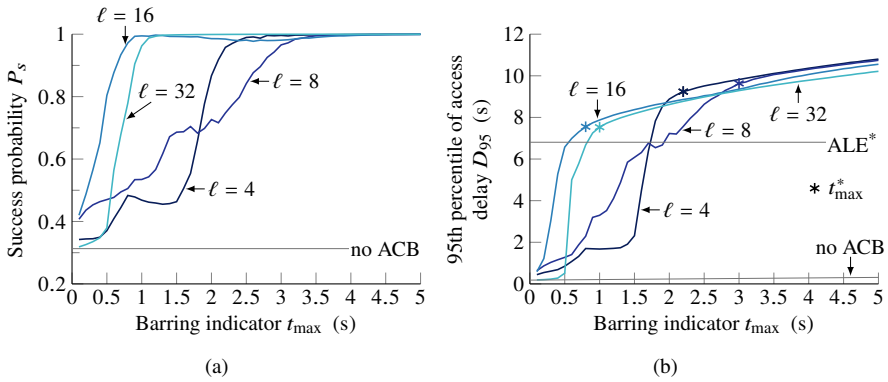


Figure C.3: (a) Success probability P_s and (b) 95th percentile of access delay D_{95} for the ACBC scheme with the RLS algorithm as a function of $t_{\max}$ under the TM 2; $r = 54$ and ω^*.

Appendix D

Publications directly related to this thesis

Journals

1. I. Leyva-Mayorga, M. A. Rodriguez-Hernandez, V. Pla, J. Martinez-Bauset and L. Tello-Oquendo, "Adaptive access class barring for efficient mMTC," *Comput. Netw.*, 2018. doi: 10.1016/j.comnet.2018.12.003.

2. L. Tello-Oquendo, V. Pla, I. Leyva-Mayorga, J. Martinez-Bauset, V. Casares-Giner, and L. Guijarro, "Efficient random access channel evaluation and load estimation in LTE-A with massive MTC," *IEEE Trans. Veh. Technol.*, 2018. doi: 10.1109/TVT.2018.2885333.

3. L. Tello-Oquendo, I. Leyva-Mayorga, V. Pla, J. Martinez-Bauset, J. R. Vidal, V. Casares-Giner, and L. Guijarro, "Performance analysis and optimal access class barring parameter configuration in LTE-A Networks with Massive M2M Traffic," *IEEE Trans. Veh. Technol.*, vol. 67, no. 4, pp. 3505–3520, 2018, doi: 10.1109/TVT.2017.2776868

4. I. Leyva-Mayorga, L. Tello-Oquendo, V. Pla, J. Martinez-Bauset, and V. Casares-Giner, "On the Accurate Performance Evaluation of the LTE-A Random Access Procedure and the Access Class Barring Scheme," *IEEE Trans. Wireless Commun.*, vol. 16, no. 12, pp. 7785–7799, 2017, doi: 10.1109/TWC.2017.2753784.

5. I. Leyva-Mayorga, V. Pla, J. Martinez-Bauset, and M. E. Rivero-Angeles, "A hybrid method for the QoS analysis and parameter optimization in time-critical random access wireless sensor networks," *J. Netw. Comput. Appl.*, vol. 83, pp. 190–203, 2017, doi: 10.1016/j.jnca.2017.01.027.

International Conferences

1. I. Leyva-Mayorga, R. Torre, S. Pandi, G. T. Nguyen, V. Pla, J. Martinez-Bauset, and F. H. P. Fitzek, "A Network-coded Cooperation Protocol for Efficient Massive Content Distribution," to be published *in Proc. IEEE Global Commun. Conf. (GLOBECOM)*, 2018.

2. I. Leyva-Mayorga, L. Tello-Oquendo, V. Pla, J. Martinez-Bauset, and V. Casares-Giner, "On the Accurate Performance Evaluation of the LTE-A Random Access Procedure," *in Proc. IEEE Global Commun. Conf. (GLOBECOM)*, 2017, pp. 1–7, doi: 10.1109/GLOCOM.2017.8254063.

3. I. Leyva-Mayorga, M. A. Rodriguez-Hernandez, V. Pla, J. Martinez-Bauset, and L. Tello-Oquendo, "An Adaptive Access Class Barring Scheme for Handling Massive M2M Commun. in LTE-A," *in Proc. Eur. Wireless Conf.,* 2017, pp. 143–148.

4. I. Leyva-Mayorga, L. Tello-Oquendo, V. Pla, J. Martinez-Bauset, and V. Casares-Giner, "Performance analysis of access class barring for handling massive M2M traffic in LTE-A networks," *in Proc. IEEE Int. Conf. Commun. (ICC)*, 2016, pp. 1–6, doi: 10.1109/ICC.2016.7510814.

5. I. Leyva-Mayorga, V. Pla and M. E. Rivero-Angeles, "A Hybrid Method for Obtaining the Distribution of Report Latency in Wireless Sensor Networks", *in Proc. IFIP Wireless and Mobile Netw. Conf. (WMNC)*, 2015, pp. 9–15. doi: 10.1109/WMNC.2015.8.

Appendix E

Research projects

This work has been developed in the framework of the following research projects:

- TIN2013-47272-C2-1-R: PLASMA -*Platform of Services for Smart Cities with Dense Machine-to-Machine Networks*.
- TEC2015-71932-REDT: Elastic Networks - *New Paradigms of Elastic Networks for a World Radically Based on Cloud and Fog Computing*.

Bibliography

[1] *Study on RAN improvements for machine-type communications*, 3GPP TR 37.868, Jul. 2011.

[2] *Architecture enhancements to facilitate communications with packet data networks and applications*, 3GPP TS 23.682 V13.5.0, Mar. 2016.

[3] *Feasibility study for further advancements for E-UTRA*, 3GPP TR 36.912 V13.0.0, Jan. 2016.

[4] *Physical layer procedures*, 3GPP TS 36.213 V13.0.0, May 2016.

[5] *Service accessibility*, 3GPP TS 22.011 V13.6.0, Jul. 2016.

[6] *Service requirements for machine-type communications*, 3GPP TS 22.368 V13.2.0, Dec. 2016.

[7] *5G; Study on scenarios and requirements for next generation access technologies*, 3GPP TR 38.913 V14.2.0, May 2017.

[8] *Physical channels and modulation*, 3GPP TS 36.211 V14.2.0, Apr. 2017.

[9] *Medium access control (MAC) protocol specification*, 3GPP TS 36.321 V15.2.0, Jul. 2018.

[10] *Radio resource control (RRC); Protocol specification*, 3GPP TS 36.331 V15.3.0, Sep. 2018.

[11] *Final Overall 5G RAN Design*, 5G-PPP METIS-II report 2.4, Jun. 2017.

[12] R. Abbas, M. Shirvanimoghaddam, Y. Li, and B. Vucetic, "Random access for M2M communications with QoS guarantees," *IEEE Trans. Commun.*, vol. 65, no. 7, pp. 2889–2903, 2017.

[13] M. M. Afsar and M.-H. Tayarani-N, "Clustering in sensor networks: a literature survey," *J. Netw. Comput. Appl.*, vol. 46, pp. 198–226, 2014.

[14] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5G wireless networks: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 1617–1655, 2016.

[15] R. Ahlswede, N. Cai, S.-Y. R. Li, and R. W. Yeung, "Network information flow," *IEEE Trans. Inf. Theory*, vol. 46, no. 4, pp. 1204–1216, 2000.

[16] I. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "Wireless sensor networks: a survey," *Comput. Netw.*, vol. 38, no. 4, pp. 393–422, 2002.

[17] A. S. Alfa, *Queueing Theory for Telecommunications*. MA, USA: Springer US, 2010.

[18] T. AlSkaif, M. Guerrero Zapata, and B. Bellalta, "Game theory for energy efficiency in wireless sensor networks: latest trends," *J. Netw. Comput. Appl.*, vol. 54, pp. 33–61, 2015.

[19] J. J. G. Andrews, S. Buzzi, W. Choi, S. V. S. Hanly, A. Lozano, A. A. C. K. Soong, and J. J. C. Zhang, "What will 5G be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, 2014.

[20] O. Arouk and A. Ksentini, "General model for RACH procedure performance analysis," *IEEE Commun. Lett.*, vol. 20, no. 2, pp. 372–375, 2016.

[21] A. Asudeh, G. V. Zaruba, and S. K. Das, "A general model for MAC protocol selection in wireless sensor networks," *Ad Hoc Netw.*, vol. 36, pp. 189–202, 2016.

[22] L. Aymen, B. Ye, and T. M. T. Nguyen, "Offloading performance evaluation for network coding-based cooperative mobile video streaming," in *Proc. Int. Conf. Netw. Future (NOF)*, 2016, pp. 1–5.

[23] R. Bellman, *Dynamic Programming*, 1st ed. NJ, USA: Princeton University Press, 1957.

[24] A. Biral, M. Centenaro, A. Zanella, L. Vangelista, and M. Zorzi, "The challenges of M2M massive access in wireless cellular networks," *Digit. Commun. Networks*, vol. 1, no. 1, pp. 1–19, 2015.

[25] C. T. Calafate, C. Lino, J.-C. Cano, and P. Manzoni, "Modeling emergency events to evaluate the performance of time-critical WSNs," in *Proc. IEEE Symp. Comput. and Commun. (ISCC)*, 2010, pp. 222–228.

[26] T. Y. Chan, Y. Ren, Y. C. Tseng, and J. C. Chen, "How to reduce unexpected eMBMS session disconnection: Design and performance analysis," *IEEE Wireless Commun. Lett.*, vol. 7, no. 1, pp. 126–129, 2018.

[27] Z. Chang, S. Zhou, T. Ristaniemi, and Z. Niu, "Collaborative mobile clouds: An energy efficient paradigm for content sharing," *IEEE Wireless Commun.*, vol. 25, no. 2, pp. 186–192, 2018.

[28] D.-R. Chen, "An energy-efficient QoS routing for wireless sensor networks using self-stabilizing algorithm," *Ad Hoc Netw.*, vol. 37, pp. 240–255, 2015.

[29] M.-Y. Cheng, G.-Y. Lin, H.-Y. Wei, and A. C.-C. Hsu, "Overload control for machine-type-communications in LTE-advanced system," *IEEE Commun. Mag.*, vol. 50, no. 6, pp. 38–45, 2012.

[30] R. G. Cheng, J. Chen, D. W. Chen, and C. H. Wei, "Modeling and analysis of an extended access barring algorithm for machine-type communications in LTE-A networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 6, pp. 2956–2968, 2015.

[31] D. Chu, "Polyphase codes with good periodic correlation properties," *IEEE Trans. Inf. Theory*, vol. 18, no. 4, pp. 531–532, 1972.

[32] (2017, Feb.) Cisco visual networking index (VNI): Global mobile data traffic forecast update, 2016–2021 white paper. Cisco. Accessed: Oct. 15, 2018. [Online]. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html

[33] M. Condoluci, G. Araniti, T. Mahmoodi, and M. Dohler, "Enabling the IoT machine age with 5G: Machine-type multicast services for innovative real-time applications," *IEEE Access*, vol. 4, no. c, pp. 5555–5569, 2016.

[34] 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 Commun. Mag.*, vol. 55, no. 3, pp. 124–131, 2017.

[35] 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 Trans. Veh. Technol.*, vol. 65, no. 12, pp. 9847–9861, 2016.

[36] *Delivery of Broadcast Content over LTE Networks*, EBU TR, July 2014.

[37] L. Ferdouse, A. Anpalagan, and S. Misra, "Congestion and overload control techniques in massive M2M systems: A survey," *Trans. Emerg. Telecommun. Technol.*, vol. 28, no. 2, p. e2936, 2015.

[38] F. H. P. Fitzek, M. Katz, and Q. Zhang, "Cellular controlled short-range communication for cooperative P2P networking," *Wireless Pers. Commun.*, vol. 18, no. 1, pp. 141–155, 2009.

[39] F. H. P. Fitzek and M. D. Katz, *Mobile clouds. Exploiting distributed resources in wireless, mobile and social networks*. United Kingdom: John Wiley and Sons, Ltd, 2014.

[40] F. Ghavimi and H.-H. Chen, "M2M communications in 3GPP LTE/LTE-A networks: Architectures, service requirements, challenges, and applications," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 2, pp. 525–549, 2015.

[41] P. Guo, T. Jiang, Q. Zhang, and K. Zhang, "Sleep scheduling for critical event monitoring in wireless sensor networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 23, no. 2, pp. 345–352, 2012.

[42] Haining Shu and Qilian Liang, "Fundamental performance analysis of event detection in wireless sensor networks," in *Proc. IEEE Wireless Commun. and Netw. Conf. (WCNC)*, vol. 4, 2006, pp. 2187–2192.

[43] D. C. Harrison, W. K. G. Seah, and R. Rayudu, "Rare event detection and propagation in wireless sensor networks," *ACM Comput. Surv.*, vol. 48, no. 4, pp. 1–22, 2016.

[44] R. Harwahyu, R.-G. Cheng, and C.-H. Wei, "Investigating the performance of the random access channel in NB-IoT," in *Proc. IEEE Veh. Technol. Conf. (VTC Fall)*, 2017, pp. 1–5.

[45] S. Haykin, *Adaptive filter theory*, 4th ed. NJ, USA: Prentice Hall, 2002.

[46] W. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "An application-specific protocol architecture for wireless microsensor networks," *IEEE Trans. Wireless Commun.*, vol. 1, no. 4, pp. 660–670, 2002.

[47] T. Ho, M. Médard, J. Shi, M. Effros, and D. R. Karger, "On randomized network coding," in *Proc. Annu. Allerton Conf. Commun. Control and Comput.*, vol. 41, no. 1, 2003, pp. 11–20.

[48] L. Hu, "Distributed code assignments for CDMA packet radio networks," *IEEE/ACM Trans. Netw.*, vol. 1, no. 6, pp. 668–677, 1993.

[49] *IMT vision – Framework and overall objectives of the future development of IMT for 2020 and beyond*, ITU-R Rec. M.2083, Sep. 2015.

[50] N. Jiang, Y. Deng, M. Condoluci, W. Guo, A. Nallanathan, and M. Dohler, "RACH preamble repetition in NB-IoT network," *IEEE Commun. Lett.*, vol. 22, no. 6, pp. 1244–1247, 2018.

[51] 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 Trans. Veh. Technol.*, vol. 66, no. 9, pp. 8595–8599, 2017.

[52] A. L. Jones, I. Chatzigeorgiou, and A. Tassi, "Binary systematic network coding for progressive packet decoding," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2015, pp. 4499–4504.

[53] C. Kalalas and J. Alonso-Zarate, "Reliability analysis of the random access channel of LTE with access class barring for smart grid monitoring traffic," in *Proc. IEEE Int. Conf. Commun. (ICC) Workshops*, 2017, pp. 724–730.

[54] L. Keller, A. Le, B. Cici, H. Seferoglu, C. Fragouli, and A. Markopoulou, "MicroCast: Cooperative video streaming on smartphones," in *Proc. Int. Conf. Mobile Syst., Appl., and Services (MobiSys)*, 2012, pp. 57–70.

[55] H. Khamfroush, D. E. Lucani, P. Pahlevani, and J. Barros, "On optimal policies for network-coded cooperation: Theory and implementation," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 2, pp. 199–212, 2015.

[56] J. N. Laneman, D. N. C. Tse, and G. W. Wornell, "Cooperative diversity in wireless networks: Efficient protocols and outage behavior," *IEEE Trans. Inf. Theory*, vol. 50, no. 12, pp. 3062–3080, 2004.

[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® Technol. J.*, vol. 18, no. 1, pp. 172–193, 2014.

[58] A. Laya, L. Alonso, and J. Alonso-Zarate, "Is the random access channel of LTE and LTE-A suitable for M2M communications? A survey of alternatives," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 4–16, 2014.

[59] S. H. Lee and L. Choi, "SPEED-MAC: speedy and energy efficient data delivery MAC protocol for real-time sensor network applications," *Wireless Netw.*, vol. 21, no. 3, pp. 883–898, 2015.

[60] I. Leyva-Mayorga, L. Tello-Oquendo, V. Pla, J. Martinez-Bauset, and V. Casares-Giner, "On the accurate performance evaluation of the LTE-A random access procedure," in *IEEE Global Commun. Conf. (GLOBECOM)*, 2017, pp. 1–7.

[61] I. Leyva-Mayorga, R. Torre, S. Pandi, G. T. Nguyen, V. Pla, J. Martinez-Bauset, and F. H. P. Fitzek, "A network-coded cooperation protocol for efficient massive content distribution," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, 2018, to be published.

[62] I. Leyva-Mayorga, V. Pla, J. Martinez-Bauset, and M. E. Rivero-Angeles, "A hybrid method for the QoS analysis and parameter optimization in time-critical random access wireless sensor networks," *J. Netw. Comput. Appl.*, vol. 83, pp. 190–203, 2017.

[63] I. Leyva-Mayorga, V. Pla, and M. E. Rivero-Angeles, "A hybrid method for obtaining the distribution of report latency in wireless sensor networks," in *Proc. IFIP Wireless and Mobile Netw. Conf. (WMNC)*, 2015, pp. 9–15.

[64] I. Leyva-Mayorga, M. E. Rivero-Angeles, and C. C. Arellano, "Priority-based multi-event reporting in hybrid wireless sensor networks," in *Proc. IEEE Int. Conf. Advanced Inf. Netw. and Appl. (AINA)*, 2014, pp. 413–420.

[65] I. Leyva-Mayorga, M. E. Rivero-Angeles, C. Carreto-Arellano, and V. Pla, "QoS analysis for a nonpreemptive continuous monitoring and event-driven WSN protocol in mobile environments," *Int. J. Distrib. Sensor Netw.*, vol. 2015, pp. 1–16, 2015.

[66] I. Leyva-Mayorga, M. A. Rodriguez-Hernandez, V. Pla, J. Martinez-Bauset, and L. Tello-Oquendo, "An adaptive access class barring scheme for handling massive M2M communications in LTE-A," in *Proc. Eur. Wireless Conf.*, 2017, pp. 143–148.

[67] ——, "Adaptive access class barring for efficient mMTC," *Comput. Netw.*, 2018.

[68] I. Leyva-Mayorga, L. Tello-Oquendo, V. Pla, J. Martinez-Bauset, and V. Casares-Giner, "Performance analysis of access class barring for handling massive M2M traffic in LTE-A networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2016, pp. 1–6.

[69] ——, "On the accurate performance evaluation of the LTE-A random access procedure and the access class barring scheme," *IEEE Trans. Wireless Commun.*, vol. 16, no. 12, pp. 7785–7799, 2017.

[70] Z. Liang, S. Feng, D. Zhao, and X. S. Shen, "Delay performance analysis for supporting real-time traffic in a cognitive radio sensor network," *IEEE Trans. Wireless Commun.*, vol. 10, no. 1, pp. 325–335, 2011.

[71] T. M. Lin, C. H. Lee, J. P. Cheng, and W. T. Chen, "PRADA: Prioritized random access with dynamic access barring for MTC in 3GPP LTE-A networks," *IEEE Trans. Veh. Technol.*, vol. 63, no. 5, pp. 2467–2472, 2014.

[72] A. Lo, Y. Law, and M. Jacobsson, "A cellular-centric service architecture for machine-to-machine (M2M) communications," *IEEE Wireless Commun.*, vol. 20, no. 5, pp. 143–151, 2013.

[73] A. Manjeshwar and D. Agrawal, "APTEEN: a hybrid protocol for efficient routing and comprehensive information retrieval in wireless," in *Proc. Int. Parallel and Distrib. Process. Symp. (IPDPS)*, 2002, p. 8.

[74] M. M. Mansour, "Optimized architecture for computing Zadoff-Chu sequences with application to LTE," in *Proc. IEEE Global Telecommun. Conf. (GLOBECOM)*, 2009, pp. 1–6.

[75] A. Marco, R. Casas, J. Sevillano Ramos, V. Coarasa, A. Asensio, and M. S. Obaidat, "Synchronization of multihop wireless sensor networks at the application layer," *IEEE Wireless Commun.*, vol. 18, no. 1, pp. 82–88, 2011.

[76] Y. Mehmood, C. Görg, M. Muehleisen, and A. Timm-Giel, "Mobile M2M communication architectures, upcoming challenges, applications, and future directions," *EURASIP J. Wireless Commun. and Networking*, vol. 2015, no. 1, pp. 1–37, 2015.

[77] S. Misra, S. Singh, M. Khatua, and M. S. Obaidat, "Extracting mobility pattern from target trajectory in wireless sensor networks," *Int. J. Commun. Syst.*, vol. 28, no. 2, pp. 213–230, 2015.

[78] U. Monaco, F. Cuomo, T. Melodia, F. Ricciato, and M. Borghini, "Understanding optimal data gathering in the energy and latency domains of a wireless sensor network," *Comput. Netw.*, vol. 50, no. 18, pp. 3564–3584, 2006.

[79] P. Osti, P. Lassila, S. Aalto, A. Larmo, and T. Tirronen, "Analysis of PDCCH performance for M2M traffic in LTE," *IEEE Trans. Veh. Technol.*, vol. 63, no. 9, pp. 4357–4371, 2014.

[80] S. Pandi, R. T. Arranz, G. T. Nguyen, and F. H. P. Fitzek, "Massive video multicasting in cellular networks using network coded cooperative communication," in *Proc. IEEE Annu. Consumer Commun. Networking Conf. (CCNC)*, 2018, pp. 1–2.

[81] A. Paramanathan, M. V. Pedersen, D. E. Lucani, F. H. Fitzek, and M. Katz, "Lean and mean: network coding for commercial devices," *IEEE Wireless Commun.*, vol. 20, no. 5, pp. 54–61, 2013.

[82] M. V. Pedersen and F. H. P. Fitzek, "Mobile clouds: The new content distribution platform," *Proc. IEEE*, vol. 100, no. Special Centennial Issue, pp. 1400–1403, 2012.

[83] Ping Zhou, Honglin Hu, Haifeng Wang, and Hsiao-hwa Chen, "An efficient random access scheme for OFDMA systems with implicit message transmission," *IEEE Trans. Wireless Commun.*, vol. 7, no. 7, pp. 2790–2797, 2008.

[84] (2014, Jun.) The evolution of mobile technologies: 1G to 2G to 3G to 4G LTE. Qualcomm. Accessed: Oct. 15, 2018. [Online]. Available: https://www.qualcomm.com/documents/evolution-mobile-technologies-1g-2g-3g-4g-lte

[85] B. Rashid and M. H. Rehmani, "Applications of wireless sensor networks for urban areas: a survey," *J. Netw. Comput. Appl.*, vol. 60, pp. 192–219, 2015.

[86] T. Rault, A. Bouabdallah, and Y. Challal, "Energy efficiency in wireless sensor networks: a top-down survey," *Comput. Netw.*, vol. 67, pp. 104–122, 2014.

[87] M. D. Renzo, M. Iezzi, and F. Graziosi, "On diversity order and coding gain of multisource multirelay cooperative wireless networks with binary network coding," *IEEE Trans. Veh. Technol.*, vol. 62, no. 3, pp. 1138–1157, 2013.

[88] A. Sharif, V. Potdar, and A. Rathnayaka, "Prioritizing information for achieving QoS control in WSN," in *Proc. IEEE Int. Conf. Advanced Inf. Netw. and Appl. (AINA)*, 2010, pp. 835–842.

[89] S. Siddiqui and S. Ghani, "Analytical model for delay distribution of PRMAC," in *Proc. IEEE Int. Conf. Frontiers Inf. Technol. (FIT)*, 2013, pp. 1–6.

[90] E. Soltanmohammadi, K. Ghavami, and M. Naraghi-Pour, "A survey of traffic issues in machine-to-machine communications over LTE," *IEEE Internet Things J.*, vol. 3, no. 6, pp. 865–884, 2016.

[91] C. W. Sørensen, A. Paramanathan, J. A. Cabrera, M. V. Pedersen, D. E. Lucani, and F. H. P. Fitzek, "Leaner and meaner: Network coding in SIMD enabled commercial devices," in *Proc. IEEE Wireless Commun. and Netw. Conf. (WCNC)*, 2016, pp. 1–6.

[92] M. Souil, A. Bouabdallah, and A. Kamal, "Efficient QoS provisioning at the MAC layer in heterogeneous wireless sensor networks," *Comput. Commun.*, vol. 43, pp. 16–30, 2014.

[93] L. Sun, H. Deng, R. K. Sheshadri, W. Zheng, and D. Koutsonikolas, "Experimental evaluation of WiFi active power/energy consumption models for smartphones," *IEEE Trans. Mobile Comput.*, vol. 16, no. 1, pp. 115–129, 2017.

[94] M. Tavana, A. Rahmati, and V. Shah-Mansouri, "Congestion control with adaptive access class barring for LTE M2M overload using Kalman filters," *Comput. Netw.*, vol. 141, pp. 222–233, 2018.

[95] L. Tello-Oquendo, I. Leyva-Mayorga, V. Pla, J. Martinez-Bauset, J. R. Vidal, V. Casares-Giner, and L. Guijarro, "Performance analysis and optimal access class barring parameter configuration in LTE-A networks with massive M2M traffic," *IEEE Trans. Veh. Technol.*, vol. 67, no. 4, pp. 3505–3520, 2018.

[96] L. Tello-Oquendo, V. Pla, I. Leyva-Mayorga, V. Pla, J. Martinez-Bauset, J. R. Vidal, V. Casares-Giner, and L. Guijarro, "Efficient random access channel evaluation and load estimation in LTE-A with massive MTC," *IEEE Trans. Veh. Technol.*, 2018, accepted for publication.

[97] L. Tello-Oquendo, D. Pacheco-Paramo, V. Pla, and J. Martinez-Bauset, "Reinforcement learning-based ACB in LTE-A networks for handling massive M2M and H2H communications," in *IEEE Int. Conf. Commun. (ICC)*, vol. 2018-May, 2018, pp. 1–7.

[98] O. Tickoo and B. Sikdar, "Modeling queueing and channel access delay in unsaturated IEEE 802.11 random access MAC based wireless networks," *IEEE/ACM Trans. Netw.*, vol. 16, no. 4, pp. 878–891, 2008.

[99] M. Tömösközi, F. H. P. Fitzek, D. E. Lucani, M. V. Pedersen, P. Seeling, and P. Ekler, "On the packet delay characteristics for serially-connected links using random linear network coding with and without recoding," in *Proc. Eur. Wireless Conf.*, 2015, pp. 1–6.

[100] R. Torre, "Offloading traffic from cellular networks using network coding and cooperation," Master Thesis, TU Dresden, Dresden, Germany, May 2017.

[101] E. Tsimbalo, A. Tassi, and R. J. Piechocki, "Reliability of multicast under random linear network coding," *IEEE Trans. Commun.*, vol. 66, no. 6, pp. 2547–2559, 2018.

[102] A. Uncini, *Fundamentals of adaptive signal processing*, ser. Signals and Communication Technology. Cham, Switzerland: Springer, 2015.

[103] P. K. Verma, R. Verma, A. Prakash, A. Agrawal, K. Naik, R. Tripathi, T. Khalifa, M. Alsabaan, T. Abdelkader, and A. Abogharaf, "Machine-to-machine (M2M) communications: A survey," *J. Netw. Comput. Appl.*, vol. 66, pp. 83–105, 2016.

[104] *LTE Multimedia Broadcast Multicast Services (MBMS)*, Viavi Solutions White paper, 2015.

[105] L. Wang, Z. Yang, L. Xu, and Y. Yang, "NCVCS: Network-coding-based video conference system for mobile devices in multicast networks," *Ad Hoc Netw.*, vol. 45, pp. 13–21, 2016.

[106] Y.-P. E. Wang, X. Lin, A. Adhikary, A. Grovlen, Y. Sui, Y. Blankenship, J. Bergman, and H. S. Razaghi, "A primer on 3GPP narrowband Internet of Things," *IEEE Commun. Mag.*, vol. 55, no. 3, pp. 117–123, 2017.

[107] Y. Wang, M. C. Vuran, and S. Goddard, "Analysis of event detection delay in wireless sensor networks," in *Proc. IEEE Int. Conf. Comput. Commun. (INFOCOM)*, 2011, pp. 1296–1304.

[108] Z. Wang and V. W. S. Wong, "Optimal access class barring for stationary machine type communication devices with timing advance information," *IEEE Trans. Wireless Commun.*, vol. 14, no. 10, pp. 5374–5387, 2015.

[109] C. H. Wei, G. Bianchi, and R. G. Cheng, "Modeling and analysis of random access channels with bursty arrivals in OFDMA wireless networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 4, pp. 1940–1953, 2015.

[110] B. Widrow, J. R. Glover, J. M. McCool, J. Kaunitz, C. S. Williams, R. H. Hearn, J. R. Zeidler, J. E. Dong, and R. C. Goodlin, "Adaptive noise cancelling: Principles and applications," *Proc. IEEE*, vol. 63, no. 12, pp. 1692–1716, 1975.

[111] B. Widrow, J. M. McCool, M. G. Larimore, and C. R. Johnson, "Stationary and nonstationary learning characteristics of the LMS adaptive filter," *Proc. IEEE*, vol. 64, no. 8, pp. 1151–1162, 1976.

[112] B. Yahya and J. Ben-Othman, "Towards a classification of energy aware MAC protocols for wireless sensor networks," *Wireless Commun. Mobile Comput.*, vol. 9, no. 12, pp. 1572–1607, 2009.

[113] W. Ye, J. Heidemann, and D. Estrin, "An energy-efficient MAC protocol for wireless sensor networks," in *Proc. 21th Annu. Joint Conf. IEEE Comput. and Commun. Societies (INFOCOM)*, vol. 3, 2002, pp. 1567–1576.

[114] O. Younis and S. Fahmy, "HEED: a hybrid, energy-efficient, distributed clustering approach for ad hoc sensor networks," *IEEE Trans. Mobile Comput.*, vol. 3, no. 4, pp. 366–379, 2004.

[115] R. Zhang, J. Pan, J. Liu, and D. Xie, "A hybrid approach using mobile element and hierarchical clustering for data collection in WSNs," in *Proc. IEEE Wireless Commun. and Netw. Conf. (WCNC)*, 2015, pp. 1566–1571.

[116] Z. Zhang, H. Chao, W. Wang, and X. Li, "Performance analysis and UE-side improvement of extended access barring for machine type communications in LTE," in *Proc. IEEE Veh. Technol. Conf. (VTC Spring)*, 2014, pp. 1–5.

[117] T. Zheng, M. Gidlund, and J. Akerberg, "WirArb: a new MAC protocol for time critical industrial wireless sensor network applications," *IEEE Sensors J.*, vol. 16, no. 7, pp. 2127–2139, 2016.