Universitat Politècnica de València

Departamento de Comunicaciones

Broadband Internetworking Research Group

UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

Ph.D. Dissertation

«Design and Performance Analysis of Access Control Mechanisms for Massive Machine-to-Machine Communications in Wireless Cellular Networks»

Author: Luis Tello-Oquendo

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

Valencia
July 2018

To my wife, Karin, and
my daughter, Sammy,
who courageously
accompanied me on
this journey.

Acknowledgment

I would first like to express my wholehearted appreciation to my advisors, Prof. Vicent and Prof. Jorge, for their worthy guidance. Besides being with me during all this time, giving me strength and courage on every step, you absolutely provided me the knowledge I needed to choose the proper direction and complete my dissertation successfully. Thank you for all the opportunities I was given to conduct my research, for your active support, patience, critical reviews, and understanding. To Prof. Vicente Casares for his constant support, guidance, and constructive feedback at every stage of my graduate work; thank you for your kindness and generosity with my family. To Prof. José Ramón Vidal for helping and collaborating with me in several studies during my research. I extend my gratitude to all the Professors and members of the GIRBA Lab. and the ITACA Research Institute.

I would also like to express my most profound gratitude to Prof. Ian Akyildiz, for having given me the opportunity and the privilege of visiting his Broadband Wireless Networking (BWN) Lab., for sharing with me his research expertise and for all the hospitality, excellent support, guidance, and inspiration during my time at the BWN Lab. Thank you for bringing brightness and delight to my research life. I am also grateful to the former and current members of the BWN Lab., for their consistent encouragement and friendship. Thank you for every moment we shared and congratulations on your hard work, you are a great example of a research group and a tremendous motivation to move forward, particularly for me.

I would like to offer my special gratitude to Prof. Frank Li for his support, constructive comments, and the general enthusiasm. To Lakshmi, Indika, and Debasish for their friendliness, eagerness, kindness, and professionalism. Thank you for our pleasant and excellent cooperation.

For this dissertation I would like to thank the Reading Committee members for their time, interest, careful review, and valuable comments. Likewise to the Oral Defense Committee members.

Also, I want to express my gratitude to the former and current lab-mates for supporting me much and for constantly willing to help me. Besides deliberating over our difficulties and findings, we further encourage each other by talking about things other than just our research. To Diego, Elena, Julián for leading me in the initial stage of this journey. To Isra for collaborating with me in several studies. To Canek, Angel, Jairo for your friendship and support. My gratitude to the colleagues of other labs who have shared this stage.

Lastly, I would like to thank my relatives for all their love and encouragement. To my parents, in-laws, and brothers who support my family and me in all our pursuits. You are always there for us. And most of all to my dearest wife, Karin, and my daughter, Sammy, your endless love, support, encouragement, patient, and your trust in me fuel me to persevere in the face of difficulties. I am indebted to you so much that words can hardly describe. Thank you. Finally, to those who have touched my life in any way since I started this journey, you all know who you are, and I am genuinely grateful for all you have done.

Thank you very much, everyone!

Abstract

Nowadays, Internet of Things (IoT) is an essential technology for the upcoming generation of wireless systems. Connectivity is the foundation for IoT, and the type of access required will depend on the nature of the application. One of the leading facilitators of the IoT environment is machine-to-machine (M2M) communication, and particularly, its tremendous potential to offer ubiquitous connectivity among intelligent devices. Cellular networks are the natural choice for emerging IoT and M2M applications, mainly because of the coverage offered by the extensive deployment of existing infrastructure. In addition, cellular technologies such as LTE Advanced (LTE-A) or narrow band IoT (NB-IoT), offer many advantages such as security, management and quality of service (QoS), among others. Furthermore, the forthcoming 5th generation (5G) is being rapidly evolved with new functionality tailored to form an attractive solution for this kind of applications.

A major challenge in cellular networks that is receiving a significant amount of attention is to make the network capable of handling massive access scenarios in which myriad devices deploy M2M communications. Specifically, the signaling in the connection attempts and the limited resources of the physical channels during the random access procedure are the main concern. On the other hand, cellular systems have seen a tremendous development in recent decades; they incorporate sophisticated technology and algorithms to offer a broad range of services. The modeling and performance analysis of these large multi-service networks is also a challenging task that might require high computational effort.

To address the above challenges, in this dissertation we first concentrate on the design and performance evaluation of novel access control schemes to deal with massive M2M communications in wireless cellular networks. Then, we focus on the performance evaluation of large multi-service networks and propose a novel analytical technique that features accuracy and computational efficiency.

Our main objective is to provide solutions to ease the congestion in the radio access network or core network when massive M2M devices try to connect to the network. We consider the following two types of scenarios: (i) massive M2M devices connect directly to cellular base stations, and (ii) they form clusters and the data is forwarded to gateways that provide them with access to the infrastructure. In the first scenario, as the number of devices added to the network is constantly increasing, the network should handle the considerable increment in access requests. In that sense, access class barring (ACB) is proposed by the 3rd Generation Partnership Project (3GPP) as a practical congestion control solution in the radio access and core network. The proper tuning of the ACB parameters according to the traffic intensity is critical, but how to do so dynamically and autonomously is a challenging task that has not been specified. Thus, this dissertation contributes to the performance analysis and optimal design of novel algorithms to implement effectively this barring scheme and overcome the challenges introduced by massive M2M communications in current and forthcoming cellular networks. In the second scenario, since the heterogeneity of IoT devices and the hardware-based cellular architectures impose even greater challenges to enable flexible and efficient communication in 5G wireless systems, this dissertation also contributes to the design of software-defined gateways (SD-GWs) in a new architecture proposed for wireless software-defined networks called SoftAir. The deployment of these SD-GWs represents an alternative solution aiming at handling both a vast number of devices and the volume of data they will be pouring into the network. The SD-GW jointly optimizes cross-layer communication functionality between heterogeneous IoT devices and software-defined radio access networks.

Another contribution of this dissertation is to propose a novel technique for the performance analysis of large multi-service networks. The design of analytic models based on Markov chains is a conventional approach to model the traffic dynamics of multi-service wireless networks, and also to evaluate their performance. They provide essential insights and help to understand the complex interactions among different network components. The underlying complexity of the network, particularly concerning its size and the ample range of configuration options, makes the solution of the analytical models computationally costly. However, a typical characteristic of these networks is that they support multiple types of traffic flows operating at different time-scales. This time-scale separation can be exploited to reduce considerably the computational cost associated to determine the key performance indicators. Thus, we propose a novel analytical modeling approach based on the transient regime analysis, that we name absorbing Markov chain approximation (AMCA). For a given computational cost, AMCA finds common performance indicators with greater accuracy, when compared to the results obtained by other alternative approximate methods proposed in the literature. AMCA has been deployed with advantage to evaluate several network technologies, such as cognitive radio or integrated services networks.

Resumen

En la actualidad, la Internet de las Cosas (Internet of Things, IoT) es una tecnología esencial para la próxima generación de sistemas inalámbricos. La conectividad es la base de IoT, y el tipo de acceso requerido dependerá de la naturaleza de la aplicación. Uno de los principales facilitadores del entorno IoT es la comunicación machine-to-machine (M2M) y, en particular, su enorme potencial para ofrecer conectividad ubicua entre dispositivos inteligentes. Las redes celulares son la elección natural para las aplicaciones emergentes de IoT y M2M, principalmente por la cobertura que ofrece el amplio despliegue de infraestructura existente. Además, las tecnologías celulares, tales como LTE Advanced (LTE-A) o narrow band IoT (NB-IoT), ofrecen numerosas ventajas como seguridad, administración y calidad de servicio (quality of service, QoS), etc. Adicionalmente, la próxima generación de redes celulares (5th generation, 5G) se está desarrollando rápidamente con nuevas funcionalidades diseñadas para formar una solución atractiva para este tipo de aplicaciones.

Un desafío importante en las redes celulares que actualmente está recibiendo bastante atención es conseguir que la red sea capaz de manejar escenarios de acceso masivo en los que una gran cantidad de dispositivos utilizan comunicaciones M2M. Concretamente, la señalización en los intentos de conexión y la limitación de recursos de los canales físicos durante el procedimiento de acceso aleatorio constituyen el principal problema. Por otro lado, los sistemas celulares han experimentado un tremendo desarrollo en las últimas décadas: incorporan tecnología sofisticada y nuevos algoritmos para

ofrecer una amplia gama de servicios. El modelado y análisis del rendimiento de estas redes multiservicio es también una tarea desafiante que podría requerir un gran esfuerzo computacional.

Para abordar los desafíos anteriores, en esta tesis doctoral nos centramos en primer lugar en el diseño y la evaluación de las prestaciones de nuevos mecanismos de control de acceso para hacer frente a las comunicaciones masivas M2M en redes celulares. Posteriormente nos ocupamos también de la evaluación de prestaciones de redes multiservicio y proponemos una nueva técnica analítica que ofrece precisión y eficiencia computacional.

Nuestro principal objetivo es proporcionar soluciones para aliviar la congestión en la red de acceso radio cuando un gran número de dispositivos M2M intentan conectarse a la red. Consideramos los dos tipos de escenarios siguientes: (i) los dispositivos M2M se conectan directamente a las estaciones base celulares, y (ii) forman grupos y los datos se envían a concentradores de tráfico (gateways) que les proporcionan acceso a la infraestructura. En el primer escenario, dado que el número de dispositivos añadidos a la red aumenta continuamente, esta debería ser capaz de manejar el considerable incremento en las solicitudes de acceso. En ese sentido, el 3rd Generation Partnership Project (3GPP) ha propuesto el access class barring (ACB) como una solución práctica para el control de congestión en la red de acceso radio y la red troncal. El ajuste correcto de los parámetros de ACB de acuerdo con la intensidad del tráfico es crítico, pero cómo hacerlo de forma dinámica y autónoma es un problema complejo cuya solución no está recogida en las especificaciones del 3GPP. Esta tesis doctoral contribuye al análisis del rendimiento y al diseño de nuevos algoritmos que implementen efectivamente este mecanismo de control de congestión, y así superar los desafíos introducidos por las comunicaciones masivas M2M en las redes celulares actuales y futuras. En el segundo escenario, dado que la heterogeneidad de los dispositivos IoT y las arquitecturas celulares basadas en hardware imponen desafíos aún mayores para permitir una comunicación flexible y eficiente en los sistemas inalámbricos 5G, esta tesis doctoral también contribuye al diseño de software-defined gateways (SD-GWs) en una nueva arquitectura

propuesta para redes inalámbricas definidas por software que se denomina SoftAir. La introducción de estos SD-GWs representa una solución alternativa que permite manejar tanto un gran número de dispositivos como el volumen de datos que estarán vertiendo en la red. El SD-GW optimiza conjuntamente la funcionalidad de comunicación entre capas, entre dispositivos IoT heterogéneos y redes de acceso radio definidas por software.

Otra contribución de esta tesis doctoral es la propuesta de una técnica novedosa para el análisis de prestaciones de redes multiservicio de alta capacidad. El diseño de modelos analíticos basados en cadenas de Markov es el enfoque convencional para modelizar la dinámica del tráfico de redes inalámbricas multiservicio, y también para evaluar sus prestaciones. Estos modelos también ayudan a comprender las interacciones complejas entre las diferentes componentes de la red. La complejidad subyacente de la red, particularmente en lo que respecta a su tamaño y la amplia gama de opciones de configuración, hace que la solución de los modelos analíticos sea costosa desde un punto de vista computacional. Sin embargo, una característica habitual de estas redes es que soportan múltiples tipos de flujos de tráfico que operan a diferentes escalas de tiempo. Esta separación de escalas temporales puede aprovecharse para reducir considerablemente el coste computacional asociado a la obtención de los principales parámetros de prestaciones. Nuestra propuesta se basa en un nuevo enfoque del modelizado analítico de sistemas que operan a diferentes escalas temporales. Este enfoque utiliza el análisis del transitorio de una serie de subcadenas absorbentes y lo denominamos absorbing Markov chain approximation (AMCA). Nuestros resultados muestran que para un coste computacional dado, AMCA calcula los parámetros de prestaciones habituales de un sistema con mayor precisión, en comparación con los resultados obtenidos por otros métodos aproximados alternativos propuestos en la literatura.

Resum

En l'actualitat, la Internet de les Coses (Internet of Things, IoT) és una tecnologia essencial per a la propera generació de sistemes sense fil. La connectivitat és la base d'IoT, i el tipus d'accés requerit dependrà de la naturalesa de l'aplicació. Un dels principals facilitadors de l'entorn IoT és la comunicació machine-to-machine (M2M) i, en particular, el seu enorme potencial per oferir connectivitat ubiqua entre dispositius intel·ligents. Les xarxes mòbils són l'elecció natural per a les aplicacions emergents de IoT i M2M, principalment per la cobertura que ofereix l'ampli desplegament d'infraestructura existent. A més, les tecnologies cel·lulars, com ara LTE Advanced (LTE-A) o narrow band IoT (NB-IoT), ofereixen nombrosos avantatges com seguretat, administració i qualitat de servei (quality of service, QoS), etc. Addicionalment, la propera generació de xarxes mòbils (5th generation, 5G) s'està desenvolupant ràpidament amb noves funcionalitats dissenyades per formar una solució atractiva per a aquest tipus d'aplicacions.

Un desafiament important en les xarxes mòbils que actualment está rebent molta atenció és aconseguir que la xarxa siga capaç de gestionar escenaris d'accés massiu en què una gran quantitat de dispositius utilitzen comunicacions M2M. Concretament, la senyalització en els intents de connexió i la limitació de recursos dels canals físics durant el procediment d'accés aleatori constitueixen el principal problema. D'altra banda, els sistemes mòbils han experimentat un gran desenvolupament en les últimes dècades: incorporen tecnologia sofisticada i nous algoritmes per oferir una àmplia gamma de serveis. El modelatge i análisi del rendiment d'aquestes xarxes multiservei

és també un desafiament important que podria requerir un gran esforç computacional.

Per abordar els desafiaments anteriors, en aquesta tesi doctoral ens centrem en primer lloc en el disseny i l'avaluació de les prestacions de nous mecanismes de control d'accés per fer front a les comunicacions massives M2M en xarxes cel·lulars. Posteriorment ens ocupem també de l'avaluació de prestacions de xarxes multiservei i proposem una nova tècnica analítica que ofereix precisió i eficiència computacional.

El nostre principal objectiu és proporcionar solucions per a alleujar la congestió a la xarxa d'accés ràdio quan un gran nombre de dispositius M2M intenten connectar-se a la xarxa. Considerem els dos tipus d'escenaris següents: (i) els dispositius M2M es connecten directament a les estacions base cel·lulars, i (ii) formen grups i les dades s'envien a concentradors de trànsit (gateways) que els proporcionen accés a la infraestructura. En el primer escenari, atès que el nombre de dispositius afegits a la xarxa augmenta contínuament, aquesta hauria de ser capaç de gestionar el considerable increment en les sol·licituds d'accés. En aquest sentit, el 3rd Generation Partnership Project (3GPP) ha proposat l'access class barring (ACB) com una solució pràctica per al control de congestió a la xarxa d'accès ràdio i la xarxa troncal. L'ajust correcte dels paràmetres d'ACB d'acord amb la intensitat del trànsit és crític, però com fer-ho de forma dinàmica i autònoma és un problema complex, la solució del qual no està recollida en les especificacions del 3GPP. Aquesta tesi doctoral contribueix a l'anàlisi del rendiment i al disseny de nous algoritmes que implementen efectivament aquest mecanisme de control de congestió, i així superar els desafiaments introduïts per les comunicacions massives M2M en les xarxes mòbils actuals i futures. En el segon escenari, atès que l'heterogeneïtat dels dispositius IoT i les arquitectures cel·lulars basades en hardware imposen desafiaments encara més grans per permetre una comunicació flexible i eficient en els sistemes sense fil 5G, aquesta tesi doctoral també contribueix al disseny de software-defined gateways (SD-GWS) en una nova arquitectura proposada per a xarxes sense fils definides per programari que s'anomena SoftAir. La introducció d'aquests SD-GWS representa una

solució alternativa que permet gestionar tant un gran nombre de dispositius com el volum de dades que estaran abocant a la xarxa. El SD-GW optimitza conjuntament la funcionalitat de comunicació entre capes, entre dispositius IoT heterogenis i xarxes d'accés ràdio definides per software.

Una altra contribució d'aquesta tesi doctoral és la proposta d'una tècnica innovadora per a l'anàlisi de prestacions de xarxes multiservei d'alta capacitat. El disseny de models analítics basats en cadenes de Markov és l'enfocament convencional per modelitzar la dinàmica del trànsit de xarxes sense fils multiservei, i també per avaluar les seves prestacions. Aquests models també ajuden a comprendre les interaccions complexes entre les diferents components de les xarxa. La complexitat subjacent de la xarxa, particularment pel que fa a la seva grandària i l'àmplia gamma d'opcions de configuració, fa que la solució dels models analítics sigui costosa des d'un punt de vista computacional. No obstant això, una característica habitual d'aquestes xarxes és que suporten múltiples tipus de fluxos de trànsit que operen a diferents escales de temps. Aquesta separació d'escales temporals pot aprofitar-se per reduir considerablement el cost computacional associat a l'obtenció dels principals paràmetres de prestacions. La nostra proposta es basa en un nou enfocament del modelitzat analític de sistemes que operen a diferents escales temporals. Aquest enfocament utilitza l'anàlisi del transitori d'una sèrie de subcadenes absorbents i l'anomenem absorbing Markov chain Approximation (AMCA). Els nostres resultats mostren que per a un cost computacional donat, AMCA calcula els paràmetres de prestacions habituals d'un sistema amb més precisió, en comparació amb els resultats obtinguts per altres mètodes aproximats alternatius proposats en la literatura.

Contents

List of Acronyms

3GPP 3rd Generation Partnership Project

4G 4th generation

5G 5th generation

AC access class

ACB access class barring

AMCA absorbing Markov chain approximation

ARQ automation repeat request

BBS base band server

BER bit error rate

CC convolutional code

CN core network

CPRI common public radio interface

CRN cognitive radio network

CSMA/CA carrier sense multiple access with collision avoidance

CTMC continuous-time Markov chain

DAG directed acyclic graph

DL downlink

DODAG destination oriented DAG

EAB extended access barring

eNodeB Evolved Node B

FEC forward error correction

GQSA generalized quasi-stationary approximation

H2H human-to-human

HARQ hybrid automation repeat request

IAD iterative aggregation/disaggregation

IoT Internet of Things

ISN integrated service network

KPI key performance indicator

LoRa Long Range

LPWA low-power wide area

LTE Long Term Evolution

LTE-A LTE Advanced

LTE-A Pro LTE Advanced Pro

M2M machine-to-machine

MAC medium access control

MIB master information block

mmWave millimeter-wave

MTC machine type communication

NB-IoT narrow band IoT

NCD nearly completely decomposable

NFC near-field communication

NRT non-real-time

PC primary channel

PDF probability distribution function

PER packet error rate

PRACH physical random access channel

PU primary user

QoS quality of service

QSA quasi-stationary approximation

RACH random access channel

RAN radio access network

RAO random access opportunity

RFID radio-frequency identification

RPL routing protocol for low power and lossy networks

RRH remote radio head

RT real-time

SC secondary channel

SD-BS software-defined base station

SD-GW software-defined gateway

SD-RAN software-defined RAN

SDN software-defined networking

SIB system information block

SINR signal-to-interference-plus-noise ratio

SMAC sleep MAC

SNR signal-to-noise ratio

SU secondary user

TTI transmission time interval

UE user equipment

UL uplink

WSN wireless sensor network

List of Figures

List of Tables

Chapter 1

Introduction

The evolution of communication networks, devices, and applications have drawn a new technological age in which everything is connected. During the last years, cellular networks have witnessed a tremendous growth in the number of connected devices and carried traffic volume, mainly due to the proliferation of the Internet and the explosive growth of mobile applications and hyper connectivity by end users. This trend will continue in the coming years as wireless systems and their stronger network coverage provided by the network evolution (e.g., 4th generation (4G) and the forthcoming 5th generation (5G) wireless systems) are expected to support more contextual, immersive mobile experiences, such as live video, augmented reality and virtual reality, among others. It is predicted that the number of connected devices will increase to 30 billion by 2023 [6], and the global mobile data traffic will achieve 49 exabytes (10^{18} bytes) per month by 2021 [7].

Internet of Things (IoT) is emerging as one of the key transforming technologies to interconnect physical objects that interact with people, other physical objects, and systems to benefit society in unprecedented ways. IoT has extended the scope of wireless communication services from interpersonal communications to smart interconnection between things and between people and things, allowing wireless communication technologies to penetrate

into broader industries and fields. One of the leading facilitators of the IoT environment is machine-to-machine (M2M) communication or machine type communication (MTC) by offering ubiquitous applications and services. Unlike human-to-human (H2H) communication, distinct features of M2M traffic require specialized and inter-operable communication technologies. Cellular networks are the natural choice to satisfy these requirements and handling a significant part of this emerging traffic due to their already existing infrastructures, extensive area coverage, and high-performance capabilities. There are already M2M subscriptions supported by cellular operators worldwide, and new enhanced techniques are currently under standardization by 3rd Generation Partnership Project (3GPP). Also M2M connectivity is driving the development of 5G wireless systems to not only enable faster data access and support greater capacity, but in addition to support a wealth of new and diverse connected devices and services that comprise the IoT [8–11].

Handling the massive number of connections generated by a large number of M2M devices (UEs) is an essential challenge in cellular networks that has received a significant amount of attention nowadays. Specifically, the signaling in the connection attempts and the limited resources of the physical channels in the random access procedure are the main factors to consider. Furthermore, the ability to adapt to changing conditions while at the same time providing new services is a constant challenge that cellular network operators have to face and one that often implies new investments in infrastructure. In this dissertation, we provide solutions to overcome the above challenges that might hinder the broad horizons of M2M applications. We consider two main scenarios in this context: (i) M2M devices access directly the network through the cellular base station, and (ii) a 5G scenario in which M2M devices are aggregated in clusters and access the network through a software-defined gateway (SD-GW).

With regard to the first scenario, the access class barring (ACB) scheme is included in the radio resource control specification as a viable congestion control scheme. Forthcoming technologies like narrow band IoT (NB-IoT) and 5G New Radio also adopt such access control. ACB spreads UE accesses

over time and enables barring of UEs with certain probability and for a random time; these parameters are broadcast by the eNodeB and are applied to UEs before they perform the first access attempt. Consequently, ACB may be effective whenever the congestion occurs sparingly and during short periods of time (in the order of several seconds). This fact goes in line with the M2M bursty traffic behavior described in [5]. The proper tuning of ACB parameters according to the traffic intensity is critical; however, how to implement this scheme efficiently and tuning its configuration parameters so that they can dynamically and autonomously adapt to the network load is a challenging task that has not been specified. This dissertation aims at successfully addressing the above challenge by devising algorithms for the dynamic and efficient implementation of the ACB scheme. Our solutions use a comprehensive view of the random access procedure and enforce the fitting of the incoming traffic to the capacity of the random access channel (RACH), by tuning the ACB parameters appropriately. Furthermore, our algorithms conform with current system specifications and can be integrated as viable solutions, allowing efficient congestion control and facilitating the coexistence of H2H and massive M2M traffic.

Regarding to the second scenario, the upcoming 5G wireless system is envisioned to support a multitude of devices and applications such as smartwatches, autonomous vehicles, IoT, the tactile Internet, among others. These applications need more sophisticated networks that not only can support high throughput, but also provide low latency in data delivery, efficient energy consumption schemes, high scalability to accommodate a large number of devices, and ubiquitous connectivity for users. To adequately satisfy the aforementioned requirements, the so-called SoftAir [3] has been proposed as a unified software-defined platform for 5G systems with network management tools and customized applications of service providers or virtual network operators. Based on SoftAir, this dissertation introduces a new architecture for provisioning quality of service (QoS) aware IoT connectivity and handling the sheer number of devices in IoT. Concretely, SD-GWs are designed in SoftAir to explore the interactions between two types of networks

[i.e., IoT and software-defined RANs (SD-RANs)] and enable cross-layer solutions that simultaneously achieve optimal energy savings and throughput gain in IoT and maximum sum-rates in SD-RAN.

Since analytical models are a conventional approach to evaluate the performance of wireless networks and to understand the interactions among different components of these rather complex networks, this dissertation also aims at devising a novel technique for the performance analysis of these multi-service networks. Accurate modeling of wireless network events allows many advantages such as determining performance parameters like the blocking probability, throughput, average transfer delay, among others. The increasing complexity of wireless networks regarding size, multiple configuration possibilities, and the interactions among different types of traffic flows makes modeling more challenging. When the wireless network models are based on continuous-time Markov chain (CTMC), we usually encounter two main common characteristics. First, the cardinality of the state-space of the CTMC is large. Second, the multiple types of traffic flows evolve at different time-scales. While the first characteristic usually makes the exact solution of the CTMC computationally intractable, the second one allows us to apply specific solution approaches that exploit the time-scale separation to reduce the computational cost. Following this approach, we propose a novel solution method named absorbing Markov chain approximation (AMCA) based on the transient regime analysis. It achieves higher accuracy for a given computational cost. Also, it applies to a wide range of time-scale separations, and therefore can be used to analytically model several applications such as cognitive radio or integrated services networks.

1.1 Dissertation objectives

The overall objectives of this Ph.D. dissertation are the following:

1. To perform an in-depth analysis of the random access procedure and the congestion control mechanisms in cellular networks aiming at optimizing their configuration parameters in massive access scenarios where devices deploy M2M communications.

2. To design and enhance the congestion control mechanisms of current and forthcoming cellular networks that allow handling massive access scenarios of M2M communications and to minimize the impact on H2H traffic.

3. To develop a unified software-defined platform for IoT networks (IoTs) and software-defined radio access networks (SD-RANs), and propose a joint optimization of protocols crossing different layers according to the devices' QoS requirements and system constraints aiming at the provision of reliable connectivity.

4. To devise an approximation method for the performance evaluation of multi-service networks from the teletraffic perspective, based on the separation of time-scales.

Analytical models have been the basis for the performance evaluation of the proposed schemes and algorithms in this dissertation. In addition to the numerical results obtained based on the developed mathematical models, extensive computer simulations coded in Matlab and C have been carried out to validate the obtained results and to assess the impact of the assumptions used by the analytical models. To overcome computational limitations, we developed suitable models or resort to approximation methods when it was necessary. Furthermore, `mex` files with a `gateway` function were used to compile and link one or more C source files so that they become callable from Matlab to diminish execution times.

The above objectives have been achieved in the framework of the following research projects: “Cooperation and Opportunism in Heterogeneous Wireless Access Networks (TIN2010-21378-C02-02),” “Platform of Services for Smart Cities with Dense M2M Networks (TIN2013-47272-C2-1-R),” and “New Paradigms of Elastic Networks for a World Radically Based on Cloud and Fog Computing (TEC2015-71932-REDT).”

1.2 Dissertation structure

The content of this dissertation has been addressed in six chapters, of which the first chapter is introductory, the last chapter concludes the dissertation and the rest of chapters have adequately been organized to detail the contributions embraced by the objectives mentioned above.

In Chapter 2, we conduct a thorough performance analysis of both the LTE Advanced (LTE-A) random access channel and the static ACB congestion control as defined in the 3GPP specifications. Specifically, we seek to enhance the performance of LTE-A in massive M2M scenarios by modifying certain configuration parameters and by the implementation of ACB. We observed that ACB is appropriate for handling sporadic periods of congestion. Then, we find the optimal configuration for its parameters so that a given QoS, in terms of successful access probability, is satisfied.

In Chapter 3, we deal with optimizing the ACB access control mechanism and proposing novel algorithms for its dynamic operation since there is no specification regarding how to dynamically adapt the barring rate parameter to the traffic load so that network overload is avoided and a better QoS can be offered. Specifically, two approaches are proposed to adapt the barring rate to different traffic conditions: the first one is based on reinforcement learning and the second one is based on the estimation of the current number of devices in backoff state. We perform extensive simulations considering several scenarios with different degrees of traffic load to validate our proposed solutions and the results show that they are useful for the efficient provisioning

of simultaneous H2H/M2M communications, especially in highly changing scenarios with bursty traffic as it can occur when M2M communications are involved.

In Chapter 4, the so-called SoftAir architecture on wireless software-defined networking is introduced to support a unified software-defined platform for QoS aware IoT systems. Then, SD-GWs are designed in SoftAir to explore the interactions between two types of networks (i.e., IoTs and SD-RANs) and enable cross-layer solutions that simultaneously achieve optimal energy savings and throughput gain in IoTs and maximum sum-rates in SD-RANs. Simulation results validate that our SoftAir solutions surpass classical IoT schemes by jointly optimizing communication functionalities for both IoTs and SD-RANs and offer reliable 5G IoT communication.

In Chapter 5, a new approximate method named AMCA is developed. It is useful for the analysis of systems that give rise to models with a large number of states and in which some elements act at different time scales. We compare the AMCA performance to that obtained by classical methods and by a recently proposed approach that aims at generalizing the conventional quasi-stationary approximation. We find that AMCA has a more predictable behavior, applies to a broader range of time-scale separations, and achieves higher accuracy for a given computational cost.

Finally, Chapter 6 summarizes the conducted work in this dissertation. The main contributions and conclusions are highlighted and the possible future lines of work are suggested.

Every chapter opens with its own introduction section, providing more details about its overall content.

Chapter 2

Performance Analysis of the Random Access Channel and Optimal Access Class Barring Parameter Configuration

2.1 Introduction

The world is moving beyond standalone devices into a new technological age in which everything is connected. Machine-to-machine (M2M) communication stands for the ubiquitous and automated exchange of information between devices on the edge of networks such as mobile devices, computers, sensors, actuators or cars inside a common network, the so-called Internet of Things (IoT). Recognizing the value of the IoT to the industry and the benefits this technological innovation brings to the public, enormous efforts are being made towards its standardization, which includes the development of projects and the organization of events that are directly related to create the environment needed for a vibrant IoT [12]. In coming years, a massive number of interconnected devices will provide ubiquitous access to information and services [6, 7]. These devices, known as user equipments (UEs), are set to exchange information autonomously in M2M applications

such as smart metering, e-healthcare, smart transportation, environmental monitoring, among others [11, 13, 14]. In those scenarios, the network congestion is expected to occur over time whenever a bulk of UEs transmit in a highly synchronized manner. There is a growing consensus that cellular networks are the best option for UE interconnection, as they provide ubiquitous coverage thanks to a widely deployed infrastructure, global connectivity, high quality of service (QoS), well-developed charging and security solutions [10, 15, 16]. Nevertheless, cellular technology was developed to handle human-to-human (H2H) traffic, where few devices (compared to the billions of M2M devices expected by 2020 [7]) communicate simultaneously. Hence, severe congestion is likely to occur when a massive number of M2M devices attempt to access the base stations [known as Evolved Node Bs (eNodeBs) in LTE Advanced (LTE-A)], resulting in performance degradation for both M2M and H2H communications [17, 18].

Several studies have demonstrated that the current random access procedure deployed in LTE-A networks is not efficient enough for managing massive M2M communications because the random access channel (RACH) suffers from overload in these scenarios [19, 20]. Consequently, the access class barring (ACB) scheme has been included in the LTE-A radio resource control specification [21] as a viable congestion control scheme. In ACB, each UE may randomly delay the beginning of its random access procedure according to a barring rate and a barring time, which are parameters broadcast by the eNodeB. As a result, ACB spreads the UE accesses through time; hence, ACB may be effective whenever the congestion occurs sparingly and during short periods (in the order of few seconds). This fact goes in line with the M2M bursty traffic behavior described in [5, 22].

In this chapter, we perform a thorough performance analysis of both the LTE-A random access procedure and the ACB congestion control scheme in scenarios with a massive number of M2M UEs that attempt to access the eNodeB in a highly synchronized manner. Specifically, the main contributions of this chapter are:

1. Analysis of the steady-state capacity of the LTE-A physical RACH.

2. The identification of the combinations of RACH parameters that enhance the access success probability in scenarios with massive M2M traffic.

3. A thorough analysis of the ACB scheme for properly tuning its parameters according to the network load. We evaluate the performance of LTE-A under the ACB scheme for a wide range of barring rates and barring times. Furthermore, we identify the optimal parameter configuration of ACB for the most congested scenario suggested by the 3rd Generation Partnership Project (3GPP) [5].

4. The comparison of the key performance indicators (KPIs) obtained for two possible backoff implementations at UE side:

   (a) a uniform backoff (as stated in the LTE-A medium access control (MAC) specification [23]);

   (b) an exponential backoff, where the backoff time of each UE depends on the number of transmissions attempted previously.

5. The comparison of the access success probability obtained for two collision models for the LTE-A random access procedure:

   (a) Collision model 1: collisions occur only at the transmission of *Msg1*;

   (b) Collision model 2: collisions occur only at the transmission of *Msg3*.

   Please refer to Section 2.3 for more specific details of the random access procedure.

During this study, we follow the 3GPP recommendations, as we have identified that the behavior of ACB is often misinterpreted [24]. Specifically, we have observed that most studies analyzing the performance of ACB assume

a fixed barring time, whereas the 3GPP specifies that this parameter is selected randomly for each barring check (process in which the UE determines its barring status, please refer to Section 2.3.3 for specific details of ACB and the barring checks) [4, 21]. Hence, our study is one of the few that evaluates the ACB performance with a randomly selected barring time.

The rest of the chapter is organized as follows. In Section 2.2, we conduct a review of the literature regarding the performance analysis of LTE-A and ACB. Then, we describe the random access in LTE-A, the physical RACH capacity, and the ACB scheme in Section 2.3. The selected traffic model, the configuration parameters, and the performance metrics for the RACH evaluation are presented in Section 2.4. Our most relevant results including the performance analysis of LTE-A and ACB are shown in Sections 2.5 and 2.6, respectively. Finally, we present our conclusions in Section 2.7.

2.2 Motivation and related work

The complexity of the random access procedure and the wide variety of configuration parameters make it challenging to evaluate the performance of LTE-A under M2M traffic. For instance, in the contention-based random access procedure (detailed in Section 2.3.1) there is no consensus regarding the moment of the four-message handshake (illustrated in Fig. 2.2) in which collisions occur. It is oftentimes assumed that all the collisions occur at the first step of the message handshake [25–27] (at the transmission of *Msg1* as suggested by the 3GPP in [5]). But more detailed studies such as [28–30], also assume that the collisions occur at the transmission of *Msg3* (the random access procedure will be explained in detail in Section 2.3). It is evident that the performance of LTE-A can be affected by these assumptions, but no study has yet compared them directly. However, regardless of the assumed outcome of the random access procedure, it has been demonstrated that, in its current form, it is not capable of handling massive M2M communications [5, 20, 22, 28, 29].

The 3GPP has provided a list of parameters which describe a typical configuration for the RACH and serve as initial guidelines for its performance analysis [5]. But commonly the performance of the LTE-A RACH is only evaluated with this particular configuration. Hence, the impact of parameters such as the backoff time of UEs and the maximum number of preamble transmissions allowed per UE on the network performance have been largely overlooked. Such is the case of [22], where a thorough mathematical analysis of the random access procedure is performed. Specifically, the authors assess the performance of LTE-A when a bulk of UEs attempt to access the eNodeB in a highly synchronized manner (as expected in most M2M applications) and obtain several KPIs specified by the 3GPP; however, only the typical RACH configuration is evaluated.

In [2], authors define the capacity of the physical random access channel (PRACH), $c(R)$, as the maximum expected number of successful UE access requests per random access opportunity (RAO), being R the number of available preambles in the system, and propose a dynamic congestion-control solution. The performance of this solution is compared with the implementation of ACB. However, since the ACB analysis is performed for a very limited selection of barring rates and barring times, the advantages of the proposed solution are magnified. Furthermore, the authors assume a constant barring time for all ACB checks, whereas the 3GPP states in [4] that the barring time is calculated randomly for each ACB check. The use of a constant barring time reduces the performance of ACB. The latter is a common problem in ACB analysis which is also present in [26], where a dynamic approach for selecting the optimal barring rate is presented. Here, authors select a constant barring time of one access opportunity, which highly differs from the protocol specification [4, 21]. Besides, it is assumed that the eNodeB is capable of updating and broadcasting the optimal barring rate at the beginning of each access opportunity, which is clearly not possible because the updating period of the system information blocks is much longer.

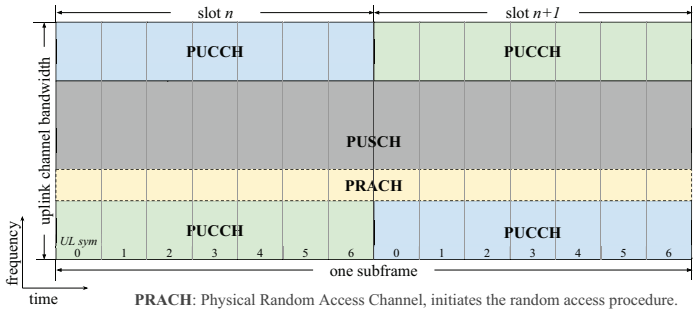
The implementation of a static barring scheme affects the access delay of every UE, even in cases of no congestion, when the scheme is not needed

at all. In these cases, the dynamic adaptation of barring parameters may be desirable, but its implementation is not straightforward. Specifically, the activation and deactivation of dynamic barring schemes are based on the collection of network congestion statistics (such as the ratio of transmitted preambles to successful accesses), which are dramatically altered whenever the barring scheme is active [26, 31]. This fact, in combination with the lack of knowledge regarding the behavior of ACB, makes it extremely tough to develop an effective adaptive ACB scheme. As such, in this study, we focus on the performance analysis of an ACB scheme whose barring parameters remain static for the entire period in which the accesses of the UEs to the eNodeB are studied. A major difference with many other studies is that we evaluate the performance of the ACB scheme considering that its parameters can take any values within the whole range suggested by the 3GPP, avoiding the restriction of these parameters only to the typical ones. This approach provides us with a wider perspective of the operation of the ACB scheme and enables the selection of optimal parameter configurations.

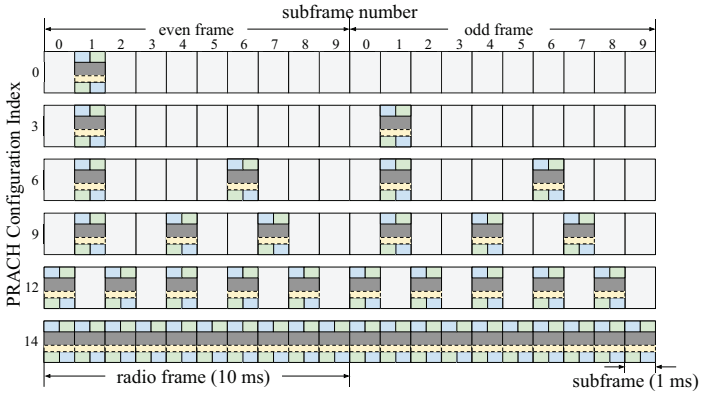
2.3 Random access in LTE-A

This section provides a general overview of the random access procedure in LTE-A networks. Two modes are defined for the random access: contention-free and contention-based. The former is used for critical situations such as handover, downlink data arrival or positioning. The latter is the standard mode for network access; it is employed by UEs to change the radio resource control state from idle to connected, to recover from a radio link failure, to perform uplink synchronization or to send scheduling requests [23].

Random access attempts of UEs are allowed in predefined time/frequency resources herein called RAOs. Two uplink channels are required; namely, the physical random access channel (PRACH) for preamble transmission and the physical uplink shared channel (PUSCH) for data transmission, see Fig. 2.1a. The PRACH is used to signal a connection request when a UE attempts to



(a)



(b)

Figure 2.1: Resource allocation in a random access cycle. (a) Physical uplink resources for initial transmission. (b) Examples of six PRACH configurations, determined by *prach-ConfigIndex*; frame structure type 1 [1].

access the cellular network. In the frequency domain, the PRACH is designed to fit in the same bandwidth as six resource blocks of normal uplink transmission ($6 \times 180\,\text{kHz}$); this fact makes it easy to schedule gaps in normal uplink transmission to allow for RAOs. In the time domain, the periodicity of the RAOs is determined by the parameter *prach-ConfigIndex*, provided by the eNodeB; a total of 64 PRACH configurations are available, Fig. 2.1b illustrates

some examples [1]. Thus, the periodicity of the RAOs ranges from a minimum of 1 RAO every two frames to a maximum of 1 RAO every subframe, i.e., from 1 RAO every 20 ms to 1 RAO every 1 ms [20, 32], [1, 21].

As mentioned before, the PRACH carries a preamble (signature) for initial access to the network; up to $R = 64$ orthogonal preambles are available per cell. In the contention-free mode, collision is avoided through the coordinated assignment of preambles, but eNodeBs can only assign these preambles during specific slots to specific UEs. In the contention-based mode, preambles are selected in a random fashion by the UEs, so there is a risk of collision, i.e., multiple UEs in the cell might pick the same preamble signature in the same RAO; therefore, contention resolution is needed. In the sequel, we focus on the analysis of the contention-based random access procedure.

2.3.1 Contention-based random access procedure

Before initiating the random access procedure, the UEs must first obtain some basic configuration parameters such as the slots in which the transmission of preambles is allowed (RAOs). The eNodeB broadcasts this information periodically through *Master Information Block* (*MIB*) and *System Information Blocks* (*SIBs*). Once the UE has acquired this information, it may proceed with the four-message handshake illustrated in Fig. 2.2. Next, we describe the four-message handshake of the contention-based random access and the backoff procedures. The interested reader is referred to [21, 23, 33, 34] for further details.

RACH preamble (Msg1): Whenever a UE attempts transmission, it sends a randomly chosen preamble in a RAO, *Msg1*. Due to the orthogonality of the different preambles, multiple UEs can access the eNodeB in the same RAO, using different preambles. The eNodeB can, without a doubt, decode a preamble transmitted (with sufficient power) by exactly one UE and estimate the transmission timing of the terminal. However, if two or more UEs transmit the same preamble, two outcomes are possible: in the first one, the

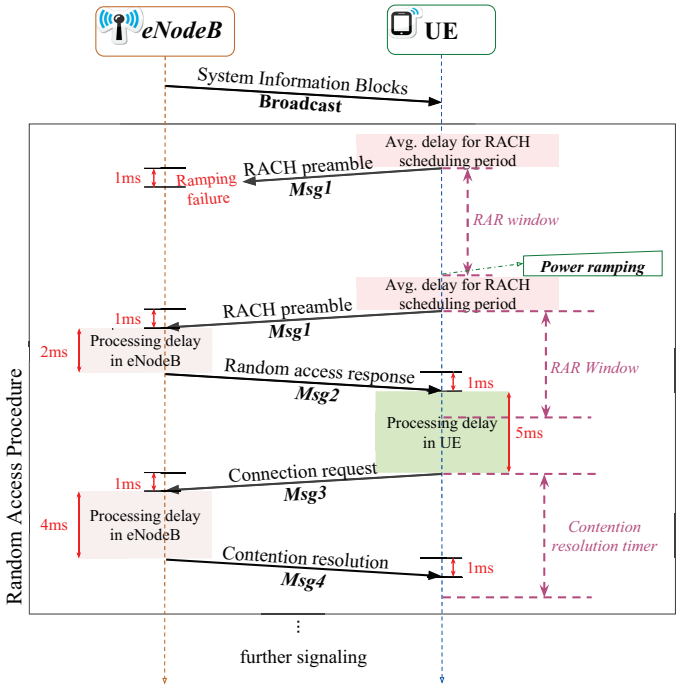


Figure 2.2: LTE-A contention-based random access procedure.

transmitted preamble cannot be decoded by the eNodeB, i.e., a collision occurs at the transmission of *Msg1* (see Fig. 2.3a) and, in the second one, the transmitted preambles are correctly decoded by the eNodeB. The main reason behind this second outcome is that the received power from one of the transmitted preambles may be much higher than the others (capture effect [35] whose quantitative evaluation is out of the scope of this dissertation); hence, the different signals may appear as a single transmission going through multiple fading paths. The preamble transmission may also fail because the UE is too far away from the eNodeB (insufficient transmission power).

Random access response—RAR—(Msg2): The eNodeB computes an identifier for each successfully decoded preamble, $\mathrm{ID} = f(\text{preamble}, \text{RAO})$,

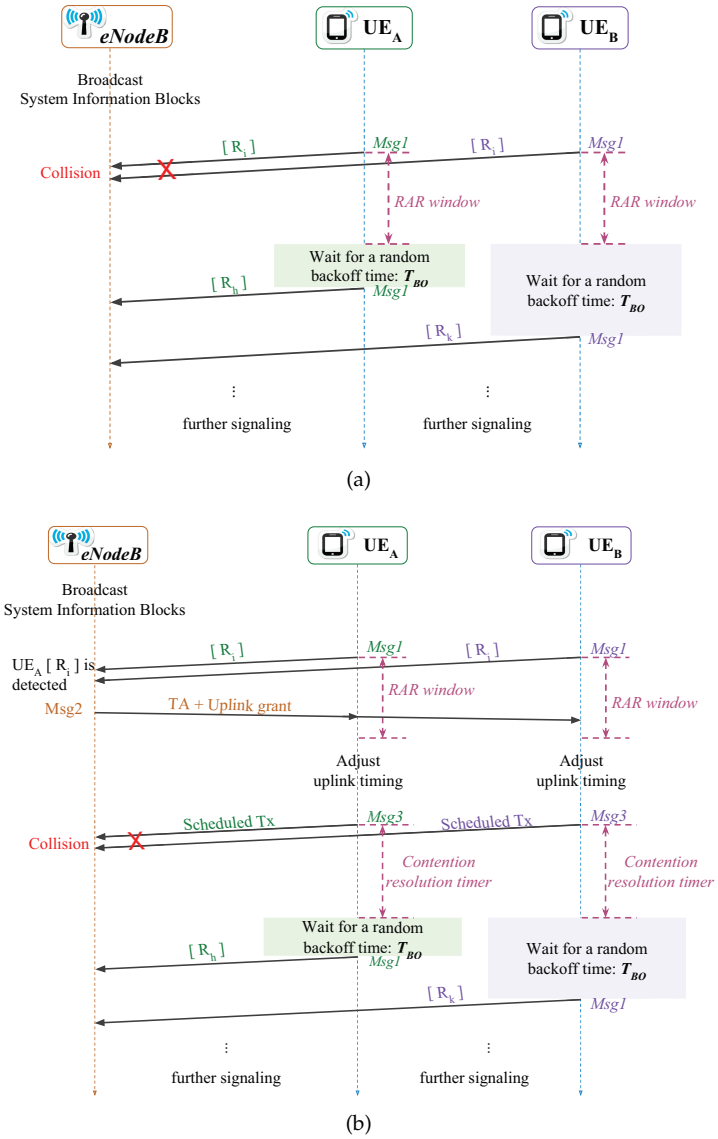


Figure 2.3: Collision outcomes in the LTE-A contention-based random access procedure. (a) Collision at the transmission of *Msg1*. (b) Collision at the transmission of *Msg3*.

and sends the RAR *Msg2* through the physical downlink control channel (PDCCH), along with the contention resolution message *Msg4*. The RAR *Msg2* includes, among other data, information about the identification of the detected preamble (ID), time alignment (TA), uplink grants (reserved PUSCH resources) for the transmission of *Msg3*, the backoff indicator (BI), and the assignment of a temporary identifier.

Exactly two subframes after the preamble transmission has ended (this is the time needed by the eNodeB to process the received preambles), the UE begins to wait for a time window, RAR window (W_{RAR}), to receive an uplink grant from the eNodeB.

There can be up to one RAR message in each subframe, but it may contain multiple uplink grants. Each uplink grant is associated to a successfully decoded preamble. The resources of the PDCCH are limited. However, the maximum number of uplink grants per RAR, N_{RAR}, can be assumed to be constant because *Msg2* (and *Msg4*) are assigned the maximum priority within the PDCCH [5]. The length of the W_{RAR} (in subframes) is broadcast by the eNodeB through the SIB Type 2 (SIB2) [21]. Hence, there is a maximum number of uplink grants that can be sent within the W_{RAR}. Only the UEs that receive an uplink grant can transmit the *Msg3*. In case the eNodeB is not capable of decoding the preambles transmitted by multiple UEs, these UEs will not receive an uplink grant (failed UEs).

Connection request (Msg3): After receiving the corresponding RAR, the UE adjusts its uplink transmission time according to the received TA and transmits a scheduled message, *Msg3*, to the eNodeB using the reserved PUSCH resources; hybrid automatic repeat request (HARQ) is used to protect the message transmission, the number of transmission attempts is limited, and the re-transmission probability for *Msg3* is set to 0.1. Recall that, if the eNodeB correctly decoded the preambles transmitted by multiple UEs, these UEs will transmit their *Msg3* over the same physical resources, thus generating a collision at this point (see Fig. 2.3b). Therefore, the eNodeB will not be able to decode the transmitted messages.

Contention resolution (Msg4): The eNodeB transmits *Msg4* as an answer to *Msg3*. The eNodeB also applies an HARQ process to send *Msg4* back to the UEs, the number of transmission attempts is limited, and the re-transmission probability for *Msg4* is set to 0.1. If a UE does not receive *Msg4* within the contention resolution timer, then it declares a failure in the contention resolution and schedules a new access attempt according the considerations detailed in the next paragraph.

If an access failure occurs at any of the steps previously described (due to insufficient transmission power or to a collision or to the expiration of the contention resolution timer), then the failed UEs ramp up their power and re-transmit a new randomly chosen preamble in a new RAO, based on a uniform backoff scheme (explained next) that uses the BI. Note that each UE keeps track of its preamble transmissions. When a UE has transmitted a certain number of preambles without success, $preambleTransMax \in \{3, 4, 5, 6, 7, 8, 10, 20, 50, 100, 200\}$ (notified by the eNodeB through a SIB), the network is declared unavailable by the UE, a random access problem is indicated to upper layers, and the random access procedure is terminated.

Backoff procedure: According to the LTE-A standard [23], if the random access attempt of a UE fails, regardless of the cause, the UE has to start the random access procedure all over again. For doing so, the UE should first perform a backoff procedure as illustrated in Fig. 2.4. In this process, the UE waits for a random time, T_{BO} [ms], until it can attempt a new preamble transmission as follows

$$T_{BO} = \mathcal{U}(0, BI), \tag{2.1}$$

where $\mathcal{U}(\cdot)$ stands for uniform distribution, BI is the backoff indicator defined by the eNodeB, and its value ranges from 0 to 960 ms. The value of BI is sent in the RAR (*Msg2*), which is read by all the UEs that sent a RACH preamble in the previous RAO. This means that every UE that did not get a *Msg2*, i.e., failed attempt, receives the BI.

Herein, we also studied the potential benefits of implementing an exponential backoff scheme, where the backoff time, T_{BO}, depends on the number

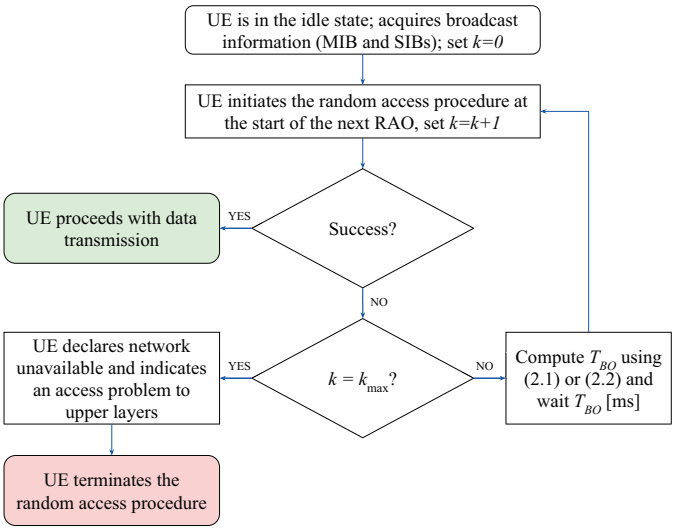


Figure 2.4: Backoff procedure performed by the failed UEs.

of preamble transmissions of each UE, $k \in \{1, 2, \ldots, k_{\max}\}$, as follows

$$T_{BO} = \mathcal{U}(0, 10 \times 2^{k-1}), \tag{2.2}$$

where the value of $k_{\max}$ is defined by the parameter *preambleTransMax*, broadcast by the eNodeB through the SIB2 [21].

2.3.2 RACH capacity

The capacity of the LTE-A RACH for the support of M2M communications is determined by two network parameters:

1) Number of available preambles: According to the LTE-A physical layer standard [1], preambles are constructed using Zadoff-Chu (ZC) sequences [36]. These sequences possess good periodic correlation properties, i.e., a negligible time is required to calculate its correlation, which allows the LTE-A system to efficiently support a large number of users per cell. Nevertheless, ZC sequences are difficult to generate in real-time due to the nature

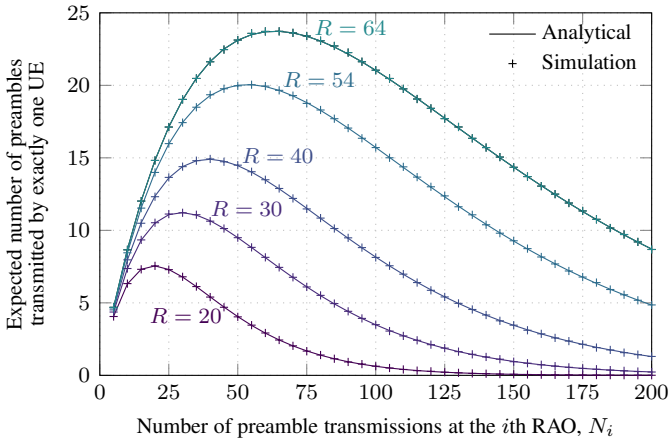


Figure 2.5: Expected number of preambles selected by exactly one UE at the ith RAO for the given number of available preambles, R, and the number of preamble transmissions, N_i [2, Fig. 3].

of their construction [37, 38] and storing them requires a significant amount of memory (around 4.9 Mbits for a pool of 64 preambles).

In [2], it is found that the capacity of the PRACH, $c(R)$, defined as the maximum expected number of preambles selected by exactly one UE in a RAO, i.e., the maximum value of the expected number of UEs that access successfully in a RAO, approximately coincides with the maximum number of stationary UE arrivals per RAO that the PRACH can handle efficiently, $\hat{c}(R)$. In other words, the performance of the PRACH drops whenever the number of UEs that begin the random access procedure at each and every RAO is $N \geq \hat{c}(R)$. If R is the number of available preambles and N_i is the number of UEs accessing at the ith RAO, it can be easily shown [2] that the expected number of preambles selected by exactly one UE is $N_i \left(1 - 1/R\right)^{N_i - 1}$ (see Fig. 2.5) and its maximum, $c(R)$, is achieved when

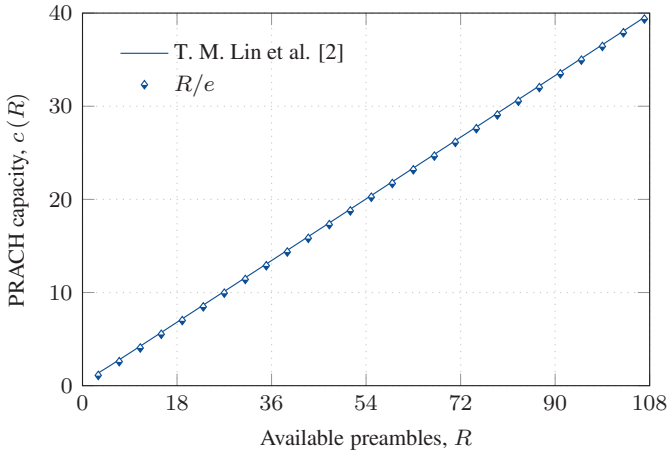


Figure 2.6: Maximum expected number of UEs that access successfully in a RAO, $c(R)$, calculated as [2] and the R/e approximation for the given number of available preambles, R.

$N_i = [\log(R/[R-1])]^{-1} \approx R$, given as follows

$$c(R) = \left[\log\left(\frac{R}{R-1}\right)\right]^{-1}\left(1-\frac{1}{R}\right)^{\left[\log\left(\frac{R}{R-1}\right)\right]^{-1}-1}, \tag{2.3}$$

which, for instance, when $R = 54$ yields $c(54) = 20.05$ successfully transmitted preambles per RAO, see Fig. 2.6. Furthermore, $c(R)$ can be approximated as follows

$$c(R) \approx R\left(1-\frac{1}{R}\right)^{R-1} \approx \frac{R}{e}. \tag{2.4}$$

The first approximation is highly accurate for practical values of R, and both of them turn out to be lower bounds of $c(R)$ as well. Please refer to Appendix B.1 for more details on this matter.

Hence, assuming a typical PRACH configuration (*prach-ConfigIndex 6*, in conformance to the LTE-A specification [5, 23]), the PRACH can handle a maximum of $\hat{c}(R) \approx 20.05$ stationary UE arrivals per RAO and, given that

RAOs occur every $T_{\text{RAO}} = 5\,\text{ms}$, a maximum of $\hat{c}(R) = \hat{c}(R)/T_{\text{RAO}} = 4010$ stationary UE arrivals per second.

2) Number of available uplink grants per RAO: Up to $N_{\text{RAR}} = 3$ uplink grants can be sent at each subframe in a RAR message, as the length of a downlink control message is 16 control channel elements (CCEs), the size of uplink grant and contention resolution messages is 4 CCEs and, at least, 4 CCEs are reserved in each subframe for a contention resolution message, *Msg4*. In the *prach-ConfigIndex 6*, RAOs occur every 5 ms (subframes) and the RAR window size (the time a UE is set to wait for the RAR) is set to $W_{\text{RAR}} = 5$ subframes. As a result, the maximum number of uplink grants that can be sent within the selected W_{RAR} is $N_{UL} = N_{\text{RAR}} \times W_{\text{RAR}} = 15$.

The performance of LTE-A plummets whenever the number of UE arrivals per RAO, N, exceeds either the PRACH capacity, $c(R)$, or the number of uplink grants that the eNodeB can send between two consecutive RAOs, N_{UL}. Thus, the main objective of congestion control schemes should be to spread UE arrivals through time to maintain the number of UE arrivals per RAO, N, below N_{UL} and $c(R)$, i.e., $N \leq \min\{N_{UL}, c(R)\}$.

2.3.3 Access class barring

ACB is a congestion control scheme designed for limiting the number of simultaneous access attempts from certain UEs according to their traffic characteristics. For doing so, all UEs are assigned to 16 mobile populations, defined as access classes (ACs) 0 to 15 (see Table 2.1). The population number is stored in UE's SIM/USIM. Each UE belongs to one out of the first 10 ACs (from ACs 0 to 9) and can also belong to one or more out of the five special categories (ACs 11 to 15). Thus, M2M devices may be assigned an AC between 0 and 9, and if a higher priority is needed, other classes may be used. In particular embodiments, AC 10 is used for an emergency call, while AC 11 to AC 15 are special high priority classes [39, 40]. Under the ACB scheme, the network operator may prevent certain UEs from making access

Table 2.1: Access Classes Defined by 3GPP [4]

Access class numbers	M2M device
0-9	Normal UEs
10	Indicates network access for Emergency Calls
11-15	Higher priority UEs

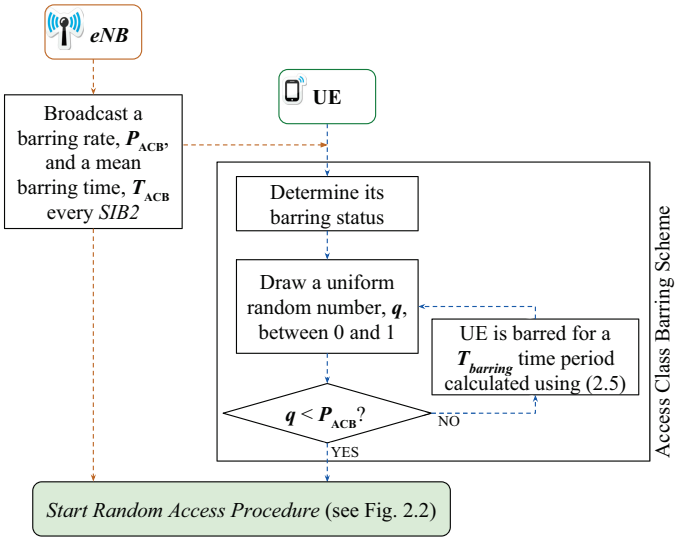


Figure 2.7: Access class barring scheme.

attempts or responding to paging messages in specific areas of a public land mobile network (PLMN) based on the corresponding AC [4, 41].

The main purpose of ACB is to redistribute the access requests of UEs through time to reduce the number of access requests per RAO. This fact helps to avoid massive-synchronized accesses demands to the PRACH, which might jeopardize the accomplishment of QoS objectives. Fig. 2.7 illustrates the ACB process [4, 21]. Note that ACB is applied to the UEs before they perform the random access procedure explained in Section 2.3.1.

If ACB is not implemented, all ACs are allowed to access the PRACH. When ACB is implemented, the eNodeB broadcasts (through SIB2) mean barring times, $T_{\text{ACB}} \in \{4, 8, 16, \ldots, 512\,\text{s}\}$, and barring rates, $P_{\text{ACB}} \in \{0.05, 0.1, \ldots$ $\ldots, 0.3, 0.4, \ldots, 0.7, 0.75, 0.8, \ldots, 0.95\}$, that are applied to ACs 0-9. Then, at the beginning of the random access procedure, each UE determines its barring status with the information provided from the eNodeB. For this, the UE generates a random number between 0 and 1, $\mathcal{U}[0, 1)$. If this number is less than or equal to P_{ACB}, the UE selects and transmits its preamble. Otherwise, the UE waits for a random time calculated as follows

$$T_{barring} = [0.7 + 0.6 \times \mathcal{U}[0, 1)] \times T_{\text{ACB}}. \tag{2.5}$$

It is worth noting that ACB is only useful for relieving sporadic periods of congestion, i.e., when a massive number of UEs attempt transmission at a given time but the system is not continuously congested.

2.4 RACH evaluation

Comparing novel congestion control schemes is not straightforward due to the large number of variables and test scenarios. For that reason, 3GPP TR 37.868 [5] has defined two different traffic models, see Table 2.2, and five KPIs to assess the efficiency of the LTE-A random access procedure with M2M communications. These directives allow for a fair comparison of novel congestion solution proposals.

Regarding the traffic models for M2M communications, traffic model 1 can be considered as a typical scenario in which the arrivals of N_M M2M UEs are uniformly distributed over a period, i.e., in a non-synchronized manner. Traffic model 2 can be seen as an extreme scenario in which a vast number of M2M UE arrivals occur in a highly synchronized manner, e.g., after an alarm that activates them.

Table 2.2: M2M Traffic Models for RACH Evaluation [5]

Characteristics	Traffic model 1	Traffic model 2
Number of M2M UEs (N_M)	1000, 3000, 5000, 10000, 30000	1000, 3000, 5000, 10000, 30000
Arrival distribution over T	Uniform	$Beta(3,4)$
Distribution period, T	60 seconds	10 seconds

2.4.1 Simulation assumptions, PRACH configuration, and performance metrics

A single cell environment is assumed to evaluate the network performance. The system accommodates both H2H and M2M UEs with different access request intensities. The access attempts of H2H UEs are distributed uniformly over time with an arrival rate of $\lambda_H = 1$ arrivals/s. Regarding the M2M UEs, $N_M = 30000$ UEs (unless otherwise stated) access the eNodeB as described in traffic model 2 (see Table 2.2). As such, we evaluate the performance of the RACH in the most congested scenario suggested by the 3GPP.

In this study, we assume a typical PRACH configuration, *prach-ConfigIndex 6*, where the subframe length is 1 ms and the periodicity of RAOs is 5 ms. $R = 54$ out of the 64 available preambles are used for contention-based random access and the maximum number of preamble transmissions of each UE, *preambleTransMax*, is set to 10. Table 2.3 lists the parameters used throughout our analysis (unless otherwise stated).

The five KPIs for the purpose of RACH capacity evaluation are the following [5]:

1. Collision probability, defined as

$$P_c = \frac{\text{Number of preambles transmitted by multiple UEs}}{R \times N_{\text{RAOs}}}, \tag{2.6}$$

Table 2.3: RACH Configuration

Parameter	Setting
PRACH Configuration Index	*prach-ConfigIndex* $= 6$
Periodicity of RAOs	5 ms
Subframe length	1 ms
Available preambles for contention-based random access	$R = 54$
Maximum number of preamble transmissions	*preambleTransMax* $= 10$
RAR window size	$W_{\text{RAR}} = 5$ subframes
Maximum number of uplink grants per subframe	$N_{\text{RAR}} = 3$
Maximum number of uplink grants per RAR window	$N_{UL} = W_{\text{RAR}} \times N_{\text{RAR}} = 15$
Preamble detection probability for the kth preamble transmission	$P_d = 1 - \frac{1}{e^k}$ [5]
Backoff Indicator	$BI = B = 20$ ms
Re-transmission probability for *Msg3* and *Msg4*	0.1
Maximum number of *Msg3* and *Msg4* transmissions	5
Preamble processing delay	2 subframes
Uplink grant processing delay	5 subframes
Connection request processing delay	4 subframes
Round-trip time (RTT) of *Msg3*	8 subframes
RTT of *Msg4*	5 subframes

where N_{RAOs} is the number of consecutive RAOs that compose the measurement period.

2. Access success probability, P_s, defined as the fraction of UEs that successfully complete the random access procedure.

3. Statistics of the number of preamble transmissions for the UEs that successfully complete the random access procedure. We assess this KPI in terms of its mean value, $\mathbb{E}[k]$.

4. Statistics of the access delay, i.e., the time elapsed between the first access attempt (preamble transmission or ACB check) and the successful completion of the random access procedure. To assess this KPI we obtain its cumulative distribution function (CDF) and the 10th, 50th and 95th percentile, D_{10}, D_{50} and D_{95}, respectively.

5. Statistics of the simultaneous preamble transmissions. We assess this KPI in terms of the maximum number of total preamble transmissions per RAO.

To obtain these KPIs, we developed two independent discrete-event simulators that allow us to corroborate our results. The first one is coded in Matlab and the second one is C-based. In each simulation, N_M UE arrivals are distributed within a period of T seconds (see Table 2.2), and the contention-based random access procedure described in Section 2.3.1 is replicated with the parameters listed in Table 2.3. Simulations are run j times until each and every one of the cumulative KPIs obtained at the jth simulation differed from those obtained at the $(j-1)$th simulation by less than 0.1 percent; different simulation seeds are used.

2.4.2 Collision Model

As mentioned in Section 2.3.1, if two or more UEs transmit the same preamble simultaneously, two outcomes are possible. In the first one, see Fig. 2.3a, a collision occurs at the transmission of *Msg1* and, in the second one, see Fig. 2.3b, a collision occurs at the transmission of *Msg3*. To evaluate the impact of these two possible outcomes on the network performance, we have defined two collision models, namely collision model 1 and collision model 2. In collision model 1, all the collisions occur at the transmission of *Msg1*, i.e.,

Table 2.4: Comparison of the Access Success Probability, P_s, for Collision Model 1 and Collision Model 2

UEs	Collision model 1	Collision model 2
M2M	31.305%	16.426%
H2H	61.335%	48.091%

the eNodeB is not capable of decoding any of the preambles transmitted by multiple UEs, so the uplink grants are only sent to the preambles transmitted by exactly one UE. In collision model 2, *Msg1* is always correctly decoded (the eNodeB successfully decodes the preambles transmitted by multiple UEs), and all the collisions occur at the transmission of *Msg3*. Note that, in practice, both types of collisions might occur. However, our interest is to study and compare the behavior of the RACH in these extreme operation scenarios. Then, the performance of real scenarios will be bounded by that of the extreme ones.

We have simulated the random access procedure with the selected traffic characteristics (traffic model 2 and $N_M = 30000$ M2M UEs) using, on the one hand, the collision model 1 and on the other hand, the collision model 2. The obtained access success probability, P_s, of both M2M and H2H UEs is shown in Table 2.4.2. It can be clearly observed that the P_s obtained under collision model 2 is much lower than the one obtained under collision model 1. This drastic reduction in P_s is mainly because in collision model 2 some uplink grants are sent in response to the transmission of a given preamble by multiple UEs, which will cause a collision during the transmission of *Msg3* and leads to (i) the waste of the limited uplink grants, and (ii) the increase of the number of contending UEs in future RAOs.

Hereafter, we select collision model 1 to conduct the performance analysis of LTE-A as suggested by the 3GPP [5] because selecting collision model 2 would magnify the increase in the performance provided by the implementation of ACB. Please note that if a different collision model is used, the

performance of the RACH will differ from the one presented in this study.

2.5 Performance analysis of LTE-A

In this section, we present some relevant results derived from our performance analysis of the LTE-A random access procedure. We begin our analysis by evaluating the capacity of the PRACH. For this, we generate a stationary distribution of $N \in \{1, 2, \ldots, 40\}$ new UE arrivals per RAO and study the effect of the number of available preambles, R, on the access success probability of UEs, P_s. To overcome the limitations of the PDCCH and evaluate the PRACH on its own, we assume that $N_{UL} = R$. Fig. 2.8 illustrates the evolution of P_s for $R \in \{20, 30, 40, 54, 64\}$. It can be observed that for each R, $P_s \approx 1$ up to a maximum value of N and then plummets. For example, when $R = 54$, $P_s \approx 1$ until $N \approx 20$, then P_s drops rapidly as N increases. Note that, $\hat{c}(R) = \{\max(N) | P_s \approx 1\}$ in a complex real scenario like the one studied is close to the PRACH capacity per RAO, $c(R)$, defined by (2.3), that was obtained using relatively simple arguments. Hence, there is a maximum stationary UE arrival rate, $\hat{c}(R) \approx c(R)$, for which UEs can efficiently access the PRACH.

Once we have studied the behavior of the PRACH in steady state, we proceed to investigate the performance of the LTE-A random access procedure according to the assumptions and the simulation parameters detailed in Section 2.4.1 and Table 2.3, respectively. As a baseline, Fig. 2.9 illustrates the expected number of UE arrivals per RAO (number of UEs that begin its random access procedure at the ith RAO), preambles with collision (collided preambles), successful accesses (UEs that complete the random access procedure successfully), and total preamble transmissions per RAO. Note that when $N_M = 30000$, traffic model 2 leads to network congestion, as the $Beta(3, 4)$ distribution of UE arrivals exceeds the PRACH capacity ($c(54) = 20.05$ UE arrivals per RAO as calculated using (2.3) and $N_{UL} = 15$) from the 343rd to the 1329th RAO. This massive number of UE arrivals results in a conges-

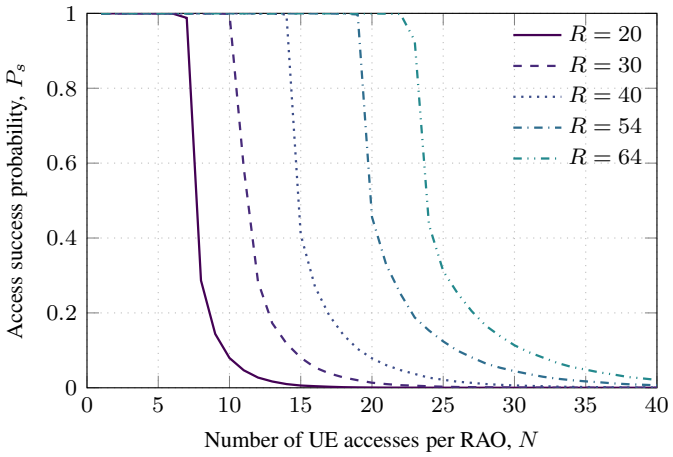


Figure 2.8: Access success probability of UEs, P_s, given the number of UE accesses per RAO, $N \in \{1, 2, \ldots, 40\}$, and the number of available preambles, $R \in \{20, 30, 40, 54, 64\}$.

tion period of $T_c = 4.93\,\text{s}$, where up to 300 average preamble transmissions per RAO occur at the 800th RAO. As a result, the average number of successful accesses sharply decreases during this period, and the access success probability is severely affected: $P_s = 31.305\,\%$.

For the remainder of this chapter, we focus on increasing the performance of the LTE-A random access procedure (assessed in terms of the KPIs defined in Section 2.4.1) when a massive number of M2M UEs, $N_M = 30000$, access the eNodeB according to traffic model 2. In the following, we investigate:

1. The number of available preambles, R, required to achieve a $P_s \approx 1$.
2. The impact of the implementation of an exponential backoff scheme instead of the standard uniform backoff scheme on the network performance.
3. The impact of the manipulation of *preambleTransMax* on the network performance.

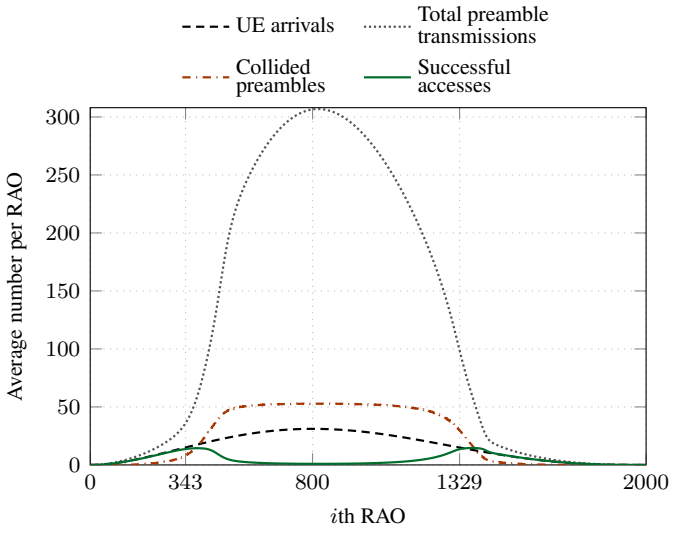


Figure 2.9: Temporal distribution of M2M UE arrivals, total preamble transmissions, collided preambles, and successful accesses; traffic model 2, $N_M = 30000$.

Next, we detail the analysis and the results of modifying the three configuration parameters mentioned above.

2.5.1 Impact of increasing the number of available preambles

To investigate whether increasing the number of available preambles can relieve congestion, we obtained the P_s of M2M UEs for several values of R, see Fig. 2.10a. Note that we assume $N_{UL} = R$. Here we observe that $P_s \geq 0.9$ is only achieved when $R \geq 90$. In other words, a dramatic increase in the number of available preambles, R, is needed to avoid PRACH congestion considering the most severely congested test scenario suggested by the 3GPP.

As mentioned in Section 2.3.2, preambles are constructed using Zadoff-

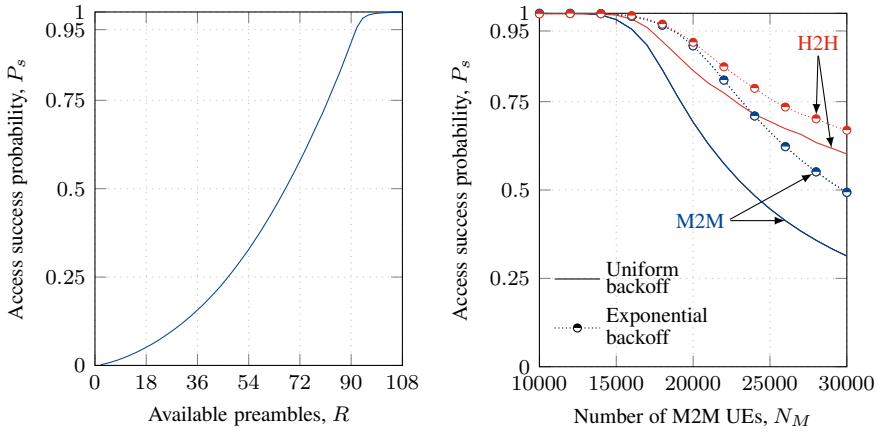


Figure 2.10: Access success probability, P_s, of M2M and H2H UEs ($\lambda_H = 1$ arrivals/s). (a) P_s of M2M UEs only given the number of available preambles, R, (b) P_s of M2M and H2H UEs given the number of M2M UEs, N_M. In (a) the number of M2M UEs is $N_M = 30000$ and the uniform backoff is used.

Chu sequences that are difficult to generate in real-time due to the nature of their construction and storing them requires a significant amount of memory. Hence, such a dramatic increase in the number of available preambles ($R \geq 90$) may not be possible. Instead of increasing the number of available preambles, R, studies such as [42–45] propose schemes for a more efficient utilization of preambles as a better solution for relieving PRACH congestion.

2.5.2 Impact of modifying the backoff scheme

According to the standard [23], UEs perform a uniform backoff, $T_{BO} = \mathcal{U}(0, BI = 20)$ ms, after a collision. We have previously observed that the use of this backoff scheme is not sufficient for relieving the congestion in the random access channels. On this basis, we investigate the use of an exponential backoff scheme, where the backoff time of each M2M UE depends on

the number of preamble transmission being attempted by that specific UE, $k \leq preambleTransMax$, and is given by (2.2). As mentioned in Section 2.4.1, the H2H UE arrivals are distributed uniformly over time, with an arrival rate of $\lambda_H = 1\,\text{arrivals/s}$.

Fig. 2.10b shows the P_s of M2M and H2H UEs when implementing the uniform and the exponential backoff schemes. On the one hand, it can be observed that the maximum number of M2M UEs that leads to $P_s \geq 0.95$ is approximately $N_M \leq 16000$ given the implementation of the uniform backoff and is $N_M \leq 19000$ when implementing the exponential backoff scheme. Hence, the use of an exponential backoff increases the number of UEs that can efficiently access the eNodeB. Nevertheless, the use of an exponential backoff is insufficient in cases of severe congestion, e.g., when $N_M \geq 20000$. On the other hand, it can also be observed from Fig. 2.10b that, in most cases, H2H UEs obtain a higher P_s than M2M UEs; this fact occurs because H2H UEs are distributed uniformly through time whereas the arrivals of M2M UEs are highly concentrate in a short time interval, i.e., between the 343rd and the 1329th RAOs. As a result, most of the H2H UEs begin its random access procedure in RAOs with a low number of preamble transmissions, where the access success probability is high. On the contrary, most of the M2M UEs begin its random access procedure in RAOs with a high number of preamble transmissions, where the access success probability is low.

2.5.3 Impact of modifying the maximum number of preamble transmissions

In Section 2.5, we have observed that severe congestion occurs when $N_M = 30000$ UEs attempt to access the eNodeB according to traffic model 2. Specifically, during the period of congestion, up to 300 preamble transmissions per RAO occur, see Fig. 2.9. Such a high number of preamble transmissions is the consequence of the fact that the higher the number of preamble transmissions in a RAO, the lower the probability of a successful preamble transmission. This fact, in turn, increases the probability of preamble re-transmissions in

the following RAOs, hence the probability of a successful preamble transmission is further reduced[1]. Therefore, during periods of congestion, the total number of preamble transmissions per RAO is highly influenced by the parameter *preambleTransMax* (maximum number of preamble transmissions). Hence, we now evaluate whether the congestion of the LTE-A random access channels can be reduced by the modification of this parameter. In Fig. 2.11 we show the P_s of M2M and H2H UEs when $preambleTransMax \in \{1, \ldots, 10\}$. Note that the highest P_s for both M2M and H2H UEs is achieved when $preambleTransMax = 3$, despite the fact that the UEs increase their transmission power at each preamble transmission, which in turn increases the preamble detection probability, P_d. These results highlight the importance of reducing congestion in order to enhance performance.

To observe more closely the behavior of LTE-A when $preambleTransMax = 3$, the average number of decoded preambles, and successful accesses per RAO given $preambleTransMax \in \{3, 10\}$ are shown in Fig. 2.12. It can be seen that a higher number of successful accesses per RAO is achieved when $preambleTransMax = 3$, which is due to a lower number of preamble transmissions per RAO. In addition to lowering congestion, which in turn increases the access success probability, reducing the number of preamble transmissions also reduces the energy consumption of UEs in highly congested scenarios. This is highly desirable because the UEs are oftentimes battery supplied.

It is worth noting that by selecting $preambleTransMax = 3$ the average number of successful accesses per RAO during congestion is close to the maximum number of uplink grants per RAO that can be sent by the eNodeB, $N_{UL} = 15$. Hence, a high percentage of the system capacity is being utilized. Nevertheless, the available uplink grants per RAO, N_{UL}, are insufficient for assigning resources to the vast number of UE arrivals. Note that combining the use of an exponential backoff with the reduction of *preambleTransMax*

[1]Please refer to Appendix B.2 in which we devise a simple closed-form expression and an efficient recursion to find the joint probability distribution function (PDF) of the number of successful and collided preambles within a RAO.

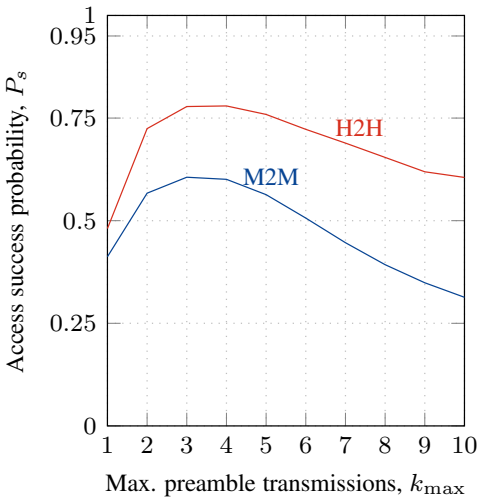


Figure 2.11: Access success probability, P_s, of M2M and H2H UEs ($\lambda_H = 1$ arrivals/s) given the maximum number of preamble transmissions, $k_{\max} =$ *preambleTransMax*.

would not be effective, i.e., in the exponential backoff, the upper limit of the backoff time increases with the number of failed preamble transmissions. Thus, the backoff time for the first few preamble transmissions is low.

In Sections 2.5.2 and 2.5.3 we have shown that either implementing an exponential backoff or reducing the maximum number of preamble transmissions increases the performance of the LTE-A RACH. However, the manipulation of neither of those parameters can prevent the system capacity from being exceeded. Yet another parameter that can be manipulated in an attempt to relieve PRACH congestion is the number of RAOs scheduled per frame. For instance, increasing the number of RAOs per frame would reduce the number of contending UEs per RAO. Nevertheless, this approach has several drawbacks: (i) it implies a reduction of the number of resources available for data transmission and, hence, a contraction of the data transport capacity of the uplink channel; (ii) the total number of RAOs that can be allocated in

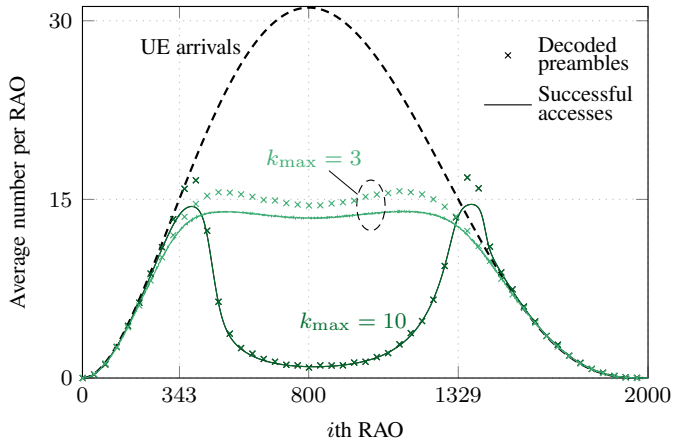


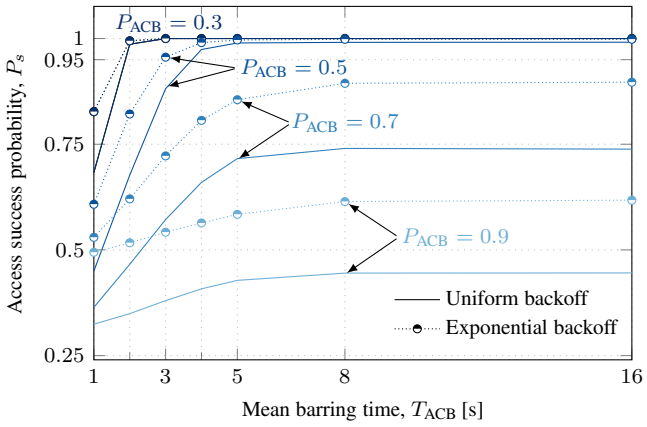
Figure 2.12: Temporal distribution of M2M UE arrivals, decoded preambles and successful UE accesses, traffic model 2, $N_M = 30000$, uniform backoff, $k_{\max} = \mathit{preambleTransMax} \in \{3, 10\}$.

an LTE-A frame is limited; and (iii) the maximum number of uplink grants that can be sent by the eNodeB per frame is fixed, so the limitations of the PDCCH remain constant.

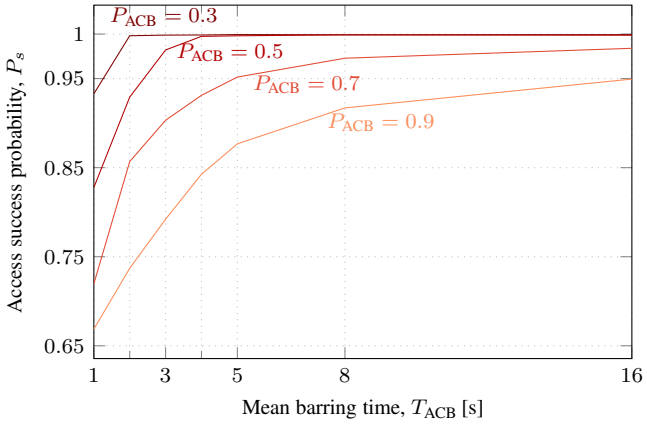
Consequently, a congestion control scheme with configurable parameters that can efficiently spread the UE arrivals through time must be implemented to drastically enhance the performance of the LTE-A. Next, we investigate the impact of the ACB congestion control scheme on the network performance.

2.6 Performance analysis of ACB

In this section, we study the impact of the implementation of the ACB scheme on the performance of LTE-A networks with massive M2M traffic. For the sake of simplicity, we assess the performance of LTE-A with an implemented ACB in terms of three KPIs, namely the access success probability, P_s, the access delay, and the average number of preamble transmissions, $\mathbb{E}[k]$, which



(a)



(b)

Figure 2.13: Access success probability of (a) M2M and (b) H2H UEs under the ACB scheme.

is closely related to energy consumption. Our main objective is to identify the configuration of ACB parameters that result in an acceptable P_s. Specifically, we aim to identify the combinations of barring rates, P_{ACB}, and barring times, T_{ACB}, that result in $P_s \geq 0.95$ for the M2M UEs.

Fig. 2.13 shows the P_s of M2M and H2H UEs, given $P_{\text{ACB}} \in \{0.3, 0.5, 0.7, 0.9\}$ and $T_{\text{ACB}} \in \{1, 2, 3, 4, 5, 8, 16\}$ s. It can be seen that, for every one of the given barring rates P_{ACB}, the access success probability, P_s, increases with the barring time, T_{ACB}. Nevertheless, for each P_{ACB} there exists a maximum value of P_s that is achieved at a certain T_{ACB}. Once this maximum P_s for each P_{ACB} is reached, further increasing T_{ACB} has no observable effect on P_s.

If we compare the P_s of the M2M UEs achieved with the implementation of a uniform backoff with the one achieved with the implementation of an exponential backoff, see Fig. 2.13a, we observe that, for the latter, shorter barring times are needed to achieve the same P_s. Please note that the H2H UEs always perform a uniform backoff. Therefore, implementing an exponential backoff in the M2M UEs does not lead to a noticeable increase in the P_s of H2H UEs, so these results have been omitted in Fig. 2.13b.

Also note that $P_s \geq 0.95$ for M2M UEs is only achieved when selecting $P_{\text{ACB}} \leq 0.5$. The effect of ACB on the UE arrivals can be closely observed in Fig. 2.14, where the average number of UE arrivals, preamble transmissions, collided preambles, and successful accesses per RAO given $P_{\text{ACB}} = 0.5$ and $T_{\text{ACB}} = 4$ s are shown. This particular combination of barring parameters leads to $P_s = 97.44\,\%$ for the M2M UEs. Such a high P_s is achieved because ACB reduces the UE arrivals per RAO from a maximum of 31.104 to 16.347, which is close to $N_{UL} = 15$ and below $\hat{c}(R) = 20.05$. As a result, we observe a dramatic reduction in the number of collisions and preamble transmissions per RAO when compared with those of Fig. 2.9.

Next, we proceed to investigate the number of preamble transmissions, k, performed by the UEs that successfully complete the random access procedure. In Fig. 2.15, we show the mean number of preamble transmissions, $\mathbb{E}[k]$, given $P_{\text{ACB}} \in \{0.3, 0.5\}$ as those barring rates lead to $P_s \geq 0.95$ (except for the lowest values of T_{ACB}, see Fig. 2.13). It can be seen that both high values of P_{ACB} and low values of T_{ACB} increase $\mathbb{E}[k]$. From Fig. 2.13 we observed that the implementation of an exponential backoff increases P_s in

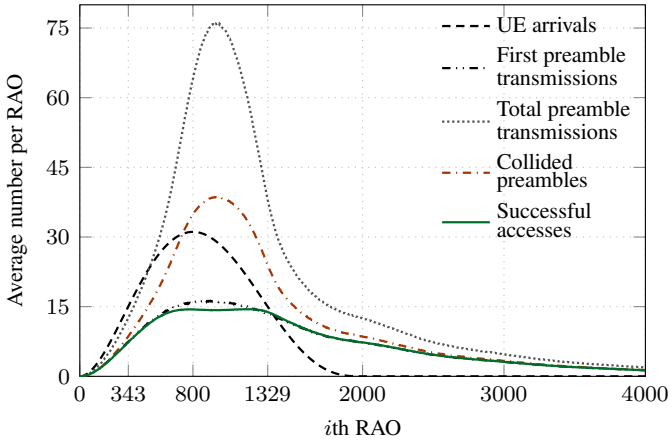


Figure 2.14: Temporal distribution of M2M UE arrivals, first preamble transmissions, total preamble transmissions, collided preambles and successful accesses, given $P_{\text{ACB}} = 0.5$ and $T_{\text{ACB}} = 4\,\text{s}$, uniform backoff.

cases where a $P_s < 0.95$ is achieved by the use of a uniform backoff. On the other hand, from Fig. 2.15 we observe that, in the mentioned cases, $\mathbb{E}[k]$ also increases. Thus, implementing an exponential backoff scheme may slightly increase P_s at the cost of increasing the energy consumption. In cases where both backoff schemes would lead to $P_s \geq 0.95$, $\mathbb{E}[k]$ is almost identical.

Finally, we studied the access delay when ACB is implemented; Fig. 2.16 illustrates these results. We calculate the access delay as the time elapsed between the arrival of a UE and the successful completion of its random access procedure, according to the timing values illustrated in Fig. 2.2 [34, Table 16.2.1-1]. For the sake of simplicity, we evaluate the access delay in terms of percentiles, defined as the maximum delay experienced by the δN UEs with the lowest delay, for the given $\delta \in \{0.1, 0.5, 0.95\}$. It is worth noting that evaluating delay in terms of its maximum achievable value, i.e., the maximum time needed for a UE to successfully complete its random access procedure, is not viable when performing ACB because this value is not upper bounded. In other words, there is no upper limit for the number of ACB checks to be

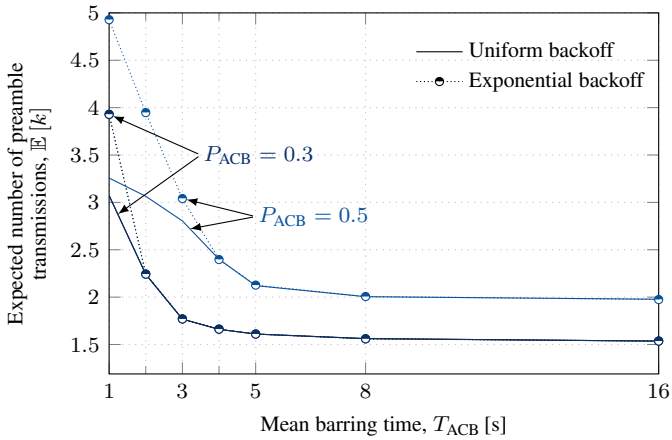
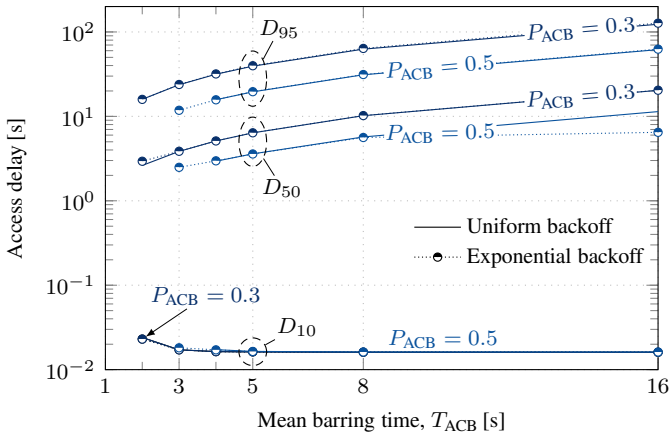


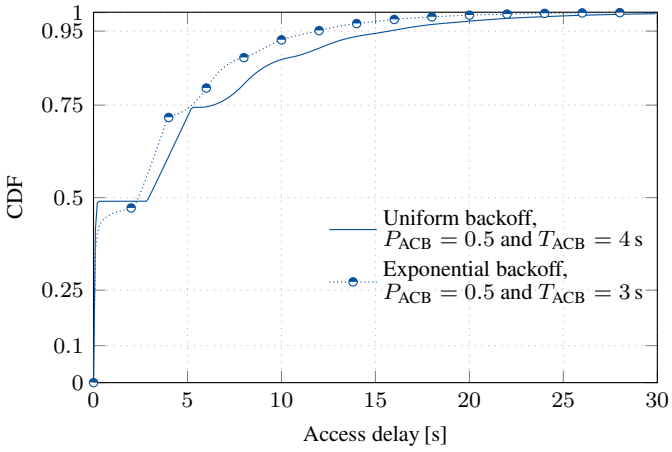
Figure 2.15: Mean number of preamble transmissions for the successfully accessed M2M UEs under the ACB scheme.

performed by a UE, hence the maximum delay, $\lim_{N_M \to \infty} D_{100} = \infty$. As such, Fig. 2.16a illustrates the 10th percentile, D_{10}, the 50th percentile, D_{50}, and the 95th percentile, D_{95}, given that $P_s \geq 0.95$.

As expected, a combination of low values of T_{ACB} with high P_{ACB} reduces the access delay. Also, though selecting a long T_{ACB} does not greatly affect P_s, see Fig. 2.13a, it sharply increases the access delay, as shown in the y-axis of Fig. 2.16a in logarithmic scale. Hence, a long T_{ACB} should be avoided. In Fig. 2.16a it can also be seen that, for the cases of interest, the delay experienced by the M2M UEs is almost the same when either a uniform or an exponential backoff scheme is implemented. Also, note that the combination of $T_{\text{ACB}} = 3\,\text{s}$ and $P_{\text{ACB}} = 0.5$ with the exponential backoff leads to $P_s \geq 0.95$. It is in this case that the overall shortest D_{50} and D_{95} are achieved. On the other hand, the shortest delay percentiles for the uniform backoff are achieved by the selection of $T_{\text{ACB}} = 4\,\text{s}$ and $P_{\text{ACB}} = 0.5$. Note that shorter delay percentiles are obtained by selecting $T_{\text{ACB}} = 3\,\text{s}$ and $P_{\text{ACB}} = 0.5$ with the uniform backoff; however, the desired $P_s \geq 0.95$ is not met as can be seen in Fig. 2.13a. It is worth mentioning that the effect of ACB in the access delay



(a)



(b)

Figure 2.16: (a) Percentiles of access delay of M2M UEs under the ACB scheme, in logarithmic scale, for the combinations of P_{ACB} and T_{ACB} that result in $P_s \geq 0.95$. (b) Cumulative distribution function of access delay for the combinations that lead to the shortest D_{50} and D_{95}, given $P_s \geq 0.95$.

Table 2.5: Access Delay of H2H UEs under the ACB Scheme

P_{ACB}	T_{ACB} [s]	D_{10} [ms]	D_{50} [ms]	D_{95} [ms]
0.3	2	15.195	20.385	56.823
	3	15.203	20.109	55.171
	4	15.202	18.286	51.187
	5	15.204	18.865	51.951
	8	15.198	16.536	50.785
	16	15.197	15.997	50.730
0.5	2	15.160	20.369	61.278
	3	15.183	20.321	60.685
	4	15.195	20.323	60.235
	5	15.193	19.304	55.264
	8	15.196	19.424	54.080
	16	15.196	17.827	50.915

of H2H UEs is almost negligible, as can be seen in Table 2.5 for $P_{\text{ACB}} \in \{3, 5\}$.

In Fig. 2.16b we compare the CDF of access delay between the selection of a uniform backoff along with $T_{\text{ACB}} = 4\,\text{s}$, $P_{\text{ACB}} = 0.5$ with that of an exponential backoff along with $T_{\text{ACB}} = 3\,\text{s}$, $P_{\text{ACB}} = 0.5$. In the former, the initial growth is much more rapid. Nevertheless, in the latter, shorter D_{50} and D_{95} are achieved.

2.6.1 Optimal ACB parameter configuration

In this section, we evaluate the performance of ACB in terms of delay and expected number of preamble transmissions which impacts on energy consumption. Recall that, if a large number of devices try to access the RACH in a short period, the preamble collisions increase significantly, resulting in huge access delays. Besides, in such a congested scenario, the repeated transmission attempts increase the energy consumption of M2M devices, most of

which will be energy-constrained. To minimize the adverse effects of congestion mentioned above, the configuration parameters of ACB, P_{ACB}, and T_{ACB}, have to be adjusted adequately. Here, we determine the optimal selection of P_{ACB} and T_{ACB} among those pairs that yields an acceptable P_s for traffic model 2 and $N_M = 30000$. For doing so, we first identify the minimum value of $T_{\text{ACB}} \in \{0.05, 0.1, \ldots\}$ [s] for a given P_{ACB} that leads to an access success probability higher than 0.95, that is,

$$T^*_{\text{ACB}} = \min\{T_{\text{ACB}} \mid P_s(P_{\text{ACB}}, T_{\text{ACB}}) \geq 0.95\}, \tag{2.7}$$

then we assess the provided QoS in terms of the expected number of preamble transmissions for the successfully accessed UEs, $\mathbb{E}^*[k]$, and the 95th percentile of access delay, D^*_{95} for the given T^*_{ACB}. The obtained T^*_{ACB}, $\mathbb{E}^*[k]$, and D^*_{95} for each $P_{\text{ACB}} \in \{0.01, 0.02, \ldots, 0.99\}$ are shown in Fig. 2.17a, Fig. 2.17b, and Fig. 2.17c, respectively, with the uniform and exponential backoff. The variability in the curves is caused by the granularity of both P_{ACB} and T_{ACB}.

The results presented in Fig. 2.17a confirm that, if the exponential backoff is selected, shorter barring times are needed to achieve $P_s \geq 0.95$ when compared to those of the uniform backoff. It can also be seen that there exists a maximum P_{ACB} for each backoff scheme that can be selected in order to achieve $P_s \geq 0.95$: 0.56 for the uniform backoff and 0.64 for the exponential backoff. Hence, the exponential backoff increases the range of P_{ACB} (and also that of T_{ACB}) that can be selected to achieve an acceptable P_s.

If we compare the average number of preamble transmissions, $\mathbb{E}^*[k]$ (see Fig. 2.17b), with the 95th percentile of access delay, D^*_{95} (see Fig. 2.17c), we clearly observe the trade-off between these KPIs; i.e., the access delay is high with configurations in which a low number of preamble transmissions are performed and vice versa.

It is worth noting that selecting T^*_{ACB} when $P_{\text{ACB}} \in [0.1, 0.6]$ only causes a slight variation in both $\mathbb{E}^*[k]$ and D^*_{95}, which is highly desirable. In addition, we can observe that the implementation of the exponential backoff increases the number of preamble transmissions but reduces the access delay when compared to the implementation of the uniform backoff.

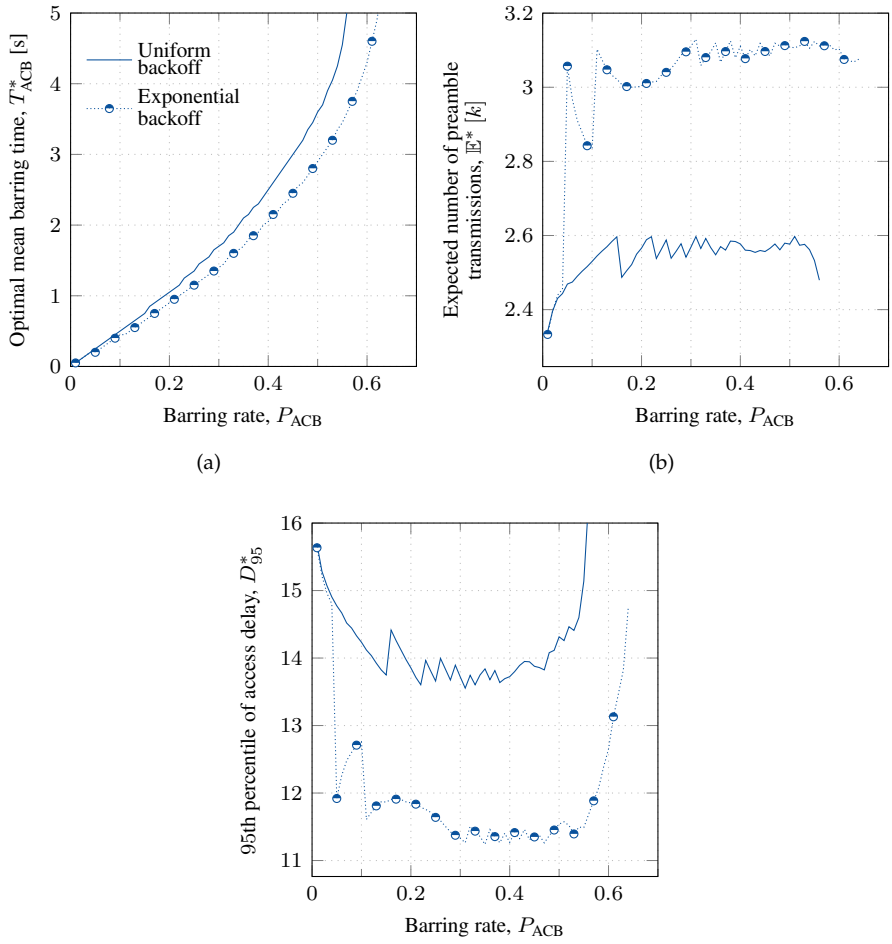


Figure 2.17: ACB optimal parameter configuration that leads to $P_s \geq 0.95$. (a) T^*_{ACB} defined as (2.7), (b) $\mathbb{E}^*[k] = \mathbb{E}[k]$ when T^*_{ACB}, and (c) $D^*_{95} = D_{95}$ when T^*_{ACB}, for the given P_{ACB}.

2.7 Highlights

We have performed a thorough study of the massive access of M2M UEs in LTE-A cellular networks. As a baseline, we obtained several KPIs to evaluate the performance of LTE-A when M2M arrivals follow either a uniform or a $Beta(3,4)$ distribution as described by traffic models 1 and 2, respectively.

We observed that traffic model 2, which describes the bursty arrivals of a massive number of M2M UEs to an eNodeB, leads to severe congestion if the eNodeB lacks a congestion control scheme. We observed that severe congestion persists regardless of the modification of network parameters such as the maximum number of allowed preamble transmissions, *preambleTransMax*, and the selected backoff scheme. Furthermore, the severity of congestion increases in cases where collisions occur during the transmissions of *Msg3*.

As such, we have studied the ACB scheme for dealing with PRACH overload and analyzed the impact of its configuration parameters on the network performance. We assume that access success probability, P_s, is the main KPI; hence, we first focus on identifying the combinations of barring rates and barring times for which the system achieves a $P_s \geq 0.95$. Then, we studied other KPIs such as the number of preamble transmissions and the access delay, where we identified a trade-off. Specifically, low barring rates and long barring times increase the access delay but reduce the number of preamble transmissions, hence reducing energy consumption.

We also compared the KPIs obtained by implementing a uniform backoff scheme, as described in the LTE-A standard [23], with that of an exponential backoff scheme along with ACB. Results show that an exponential backoff leads to a slightly higher success probability but also increases the mean number of preamble transmissions. Therefore, implementing an exponential backoff may enhance the access success probability at the cost of a higher energy consumption. Moreover, the increase in P_s provided by the exponential backoff allows the selection of lower barring times when compared to a uniform backoff. This fact, in turn, may slightly reduce the access delay.

Finally, by adequately selecting the ACB barring rates and barring times, network congestion may be relieved, even for the most congested scenario defined by the 3GPP. As such, ACB was shown to be an efficient scheme for congestion control in the RACH.

Chapter 3

Dynamic ACB Algorithms for Efficient Congestion Control

3.1 Introduction

The ability to adapt to changing conditions while at the same time providing new services is a constant challenge that cellular network operators have to face and one that very often implies new investments on infrastructure. At the same time, the high level of success of mobile technologies and their ability to easily recollect large amounts of information on users' behavior allows for a better understanding of the demand on the network and hence the provision of new solutions for the optimization of its resources. This type of approach has been used for different purposes such as access optimization and improvement of the QoS in 3G networks [46,47], or location management optimization [48], among others.

A relevant problem in cellular networks that has received an important amount of attention is the management of the massive number of connection attempts of a large number of UEs (e.g., M2M devices) since the random access channel (RACH) suffers from overload in these scenarios [19,20]. Consequently, the access class barring (ACB) scheme is included in the LTE-A Radio

Resource Control specification [21] as a viable congestion control scheme.

In Chapter 2, we conclude that there is a trade-off between relieving congestion and the key performance indicators (KPIs) of the network when the ACB is operating and its parameters are adjusted adequately [49]. Therefore, the proper tuning of ACB parameters according to the traffic intensity is critical, but the 3rd Generation Partnership Project (3GPP) does not specify any particular algorithm for that purpose.

In this chapter, we propose two algorithms for overcoming the aforementioned gap. On the one hand, a reinforcement learning (RL) approach to tune dynamically the barring rate is proposed; concretely, we use Q-learning, a well-known RL technique [50], to dynamically and autonomously tune the barring rate such that it can rapidly react to the traffic changes using local information available at the eNodeB. Our experiments for this approach are based on realistic traffic behavior by making use of traces from cellular network operators to enhance the access control of simultaneous human-to-human (H2H) and machine-to-machine (M2M) communications in LTE/LTE-A networks. On the other hand, we propose a dynamic ACB based on the estimation of UEs that are in backoff state and a mechanism for fitting this traffic jointly with the new arrivals into the RACH capacity. Thus, this scheme tunes the barring rate according to the traffic load in real-time and deactivates itself when the traffic is below the RACH capacity.

We follow the 3GPP directives [21, 23, 33] so that our proposed solutions are aligned with the network specifications and can be successfully integrated in the system. We study the behavior of our proposed algorithms in terms of key performance indicators (KPIs) suggested by the 3GPP [5] for this kind of studies. Moreover, we tested several traffic loads following the traffic model and configuration parameters for M2M communications suggested by the 3GPP [5]. Also, we evaluate the impact on the H2H traffic.

3.2 Motivation and related work

There have been a number of research efforts devoted to optimizing ACB for handling massive M2M connection attempts on the RACH through either static or dynamic approaches as detailed below; however, many of them offer complex procedures, use questionable assumptions for getting high performance, does not conform with LTE-A specifications, e.g., without considering the updating period of notification information by the evolved Node B (eNodeB), and even misinterpret the behavior of this mechanism [51–53].

In [54], a self-organizing mechanism which aims to optimize the performance of the random access procedure is proposed for M2M and H2H traffic. However, unlike the standards, the authors assume that a control-loop for congestion between the UEs and the eNodeB is available, which generates signaling load. In [52], a dynamic mechanism for access control in LTE-A is proposed to reduce the impact that massive M2M communications have on H2H traffic. Also, in this work they differentiate M2M traffic, allowing prioritization. However, this approach modifies ACB so that it can send different parameters for different classes in a similar way to extended access barring [4].

Since the number of UEs trying to access the cellular network is dynamic and this number is not known a priori, any mechanism that aims at optimizing the ACB has to develop an estimation of this value. In [26], a dynamic scheme for ACB is proposed. It is based on a Kalman filter and enhances the overall performance. Although in this work no modifications are done over the ACB mechanism, it is not possible to estimate the impact that M2M traffic has over H2H traffic, since only the first traffic type was considered. Also in [55], an optimal value of the P_{ACB} parameter is obtained in an ideal case, i.e., assuming the eNodeB has all the information about the system. Some heuristics which resemble this optimal solution are provided as well; one of them changes the parameter P_{ACB} and the other changes both P_{ACB} and the number of preambles that can be acknowledged. This solution assumes that when a UE suffers a collision, it will retry in the following RAO, which is not

consistent with the LTE-A specifications.

There have already been proposals based on reinforcement learning to optimize the access control of M2M UEs in cellular networks. In [56], the authors propose a Q-learning approach for a scenario where M2M and H2H traffic coexist. In this case, the RL scheme is performed only on the M2M UEs to allocate the random access slot on which they should transmit for avoiding collisions. Nonetheless, this scheme does not consider ACB, or the parameters that can enhance access control. In [57], the authors propose a Q-learning approach that aims at adapting the P_{ACB} as a function of the current traffic. However, they assume that the eNodeB knows the total number of contending users on each RAO to define the state space, which is not realistic. Also, they only consider a single type of traffic.

The dynamic adaptation of the barring rate is desirable since the implementation of a static ACB affects the access delay of every UE, even in cases of no congestion, when the scheme is not needed at all. To successfully address the above limitations, in Section 3.3 we present a dynamic ACB based on reinforcement learning and in Section 3.4 a dynamic ACB based on the estimation of UEs that are in backoff state. Both solutions tune the barring rate according to the traffic load in real-time and deactivates themselves when the congestion is alleviated. They conforms with current system specifications and constraints, offering effective access control.

3.3 Reinforcement learning approach

Q-learning belongs to the category of temporal-difference RL techniques that consist of learning how to map situations to actions for maximizing a scalar reward. This learning is achieved through the interaction with the environment: an agent tries an action at a specific state, and evaluates its outcomes in terms of the immediate reward or penalty it receives and its estimate of the value of the state to which it is taken. By trying all actions in all states repeatedly, it learns which are best overall, judged by long-term discounted

reward [58].

Through this approach, the eNodeB stores a value function $Q(s,a)$ that measures the expected reward from taking a given action a being on a given state s and then continuing indefinitely by taking actions optimally. In the following, we introduce the arguments and parameters that define $Q(s,a)$.

Let $\mathcal{A} = \{1,2,..,16\}$ be the set of actions that change P_{ACB} to one of its possible values defined in Section 2.3.3. When the chosen action is $a = 1$, then $P_{\text{ACB}} = 0.05$, and the rest of the values are mapped sequentially. The ACB mechanism is turned off when $P_{\text{ACB}} = 1$ (i.e., $a = 16$). Due to the characteristics of ACB, changes on P_{ACB} can only be received by UEs through SIB2 messages that are broadcast every T_{SIB2}. Hence, the Q-learning actions that change P_{ACB} will only be taken before the transmission of a SIB2. Following the specifications [21], throughout this work we will use a value of $T_{\text{SIB2}} = 16$ RAOs (i.e., 80 ms). Let s be the state defined as $s = \left(\mathbb{E}[N_{PT}], CV_{N_{PT}}, \Delta N_{PT}, P_{\text{ACB}}\right)$, where $\mathbb{E}[N_{PT}] = \frac{1}{T_{\text{SIB2}}}\left(\sum_{k=1}^{T_{\text{SIB2}}} N_{PT_k}\right)$ is the mean number of preamble transmissions per RAO that the eNodeB detected in a whole T_{SIB2} (actually a truncated version as explained below),

$$CV_{N_{PT}} = \left[\frac{1}{T_{\text{SIB2}} - 1}\left(\sum_{k=1}^{T_{\text{SIB2}}} \mid N_{PT_k} - \mathbb{E}[N_{PT}] \mid^2\right)\right]^{\frac{1}{2}} / \, \mathbb{E}[N_{PT}]$$

is the variation coefficient of N_{PT} for the same period, ΔN_{PT} is the difference between the mean number of preamble transmissions in the current period and in the previous one, and P_{ACB} is the ACB probability that affected UEs during this period.

Fig. 3.1 illustrates the above state definition. At time $t = n - 1$, which occurs just before the transmission of SIB2(n), the eNodeB decides to take an action a_n based on the state s_{n-1}. The information about the action (i.e., P_{ACB}) is sent in the following SIB2, and hence the access of UEs during the following T_{SIB2} will depend on this information. At time $t = n$, just before sending SIB2$(n+1)$, the eNodeB can calculate the values of the state s_n. For that, it will consider the 16 RAOs that lie between SIB2(n) and SIB2$(n+$

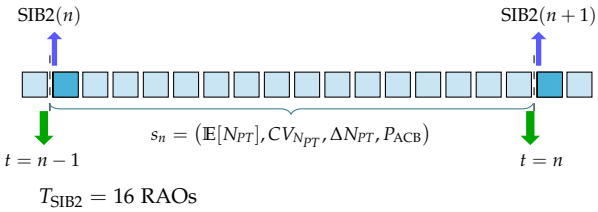


Figure 3.1: State definition and T_{SIB2}.

1). It should be noted that N_{PT} can be greater than $W_{\text{RAR}} \times N_{\text{RAR}}$, and that N_{PT} only accounts for the preamble transmissions that the eNodeB could detect properly. Hence, it is a convenient indicator of the load on the access procedure for the eNodeB.

Although there are 54 preambles available for the UEs, in our experiments we observed that even in low congested scenarios, it was very unlikely that $\mathbb{E}[N_{PT}]$ would exceeded 30. Therefore, considering that these scenarios might be related with very high congestion and that changes on P_{ACB} provide little or no improvement over the KPIs of the system, we decided to aggregate all states where $29 < \mathbb{E}[N_{PT}]$. Hence, the possible values for $\mathbb{E}[N_{PT}]$ are between 0 and 29 and we discretize $\mathbb{E}[N_{PT}]$ to ranges $0 \leq \mathbb{E}[N_{PT}] \leq 3$, $3 < \mathbb{E}[N_{PT}] < 7$, $7 \leq \mathbb{E}[N_{PT}] \leq 10$, and $10 < \mathbb{E}[N_{PT}] \leq 29$ that indicate light traffic, moderate traffic, heavy traffic, and severe traffic, respectively. On the other hand, likewise based on our observations, the coefficient of variation values $CV_{N_{PT}} \in \{0, 0.2, 0.4, 0.6, 0.8\}$ were discretized to reduce the total number of states as follows. If $0 \leq CV_{N_{PT}} \leq 0.19$ over the corresponding T_{SIB2}, then the value that will be considered to define a state will be 0; the same procedure is done for the other intervals. The parameter $\Delta N_{PT}(n)$ is obtained as $\mathbb{E}[N_{PT}](n) - \mathbb{E}[N_{PT}](n-1)$. However, this value is also discretized as follows. If $\Delta N_{PT}(n) > 0$, then it will be set as 1; if $\Delta N_{PT}(n) = 0$, then it will be set as 2; if $\Delta N_{PT}(n) < 0$, then it will be set as 3. Finally, $P_{\text{ACB}}(n)$ is the barring factor that affected UEs during the considered period, that is, the value sent in SIB2(n).

The Q-value is updated according to the Q-function [50]:

$$Q(s,a) = Q(s,a) + \alpha \left[\mathcal{R} + \gamma \max_{a' \in \mathcal{A}} \left[Q(s',a') \right] - Q(s,a) \right]; \qquad (3.1)$$

where α, $\mathcal{R}$, γ are explained as follows.

- α is the learning rate that affects how aggressive the algorithm is in adopting a new reward value into its Q-value. A higher learning rate means the algorithm will adapt to a new environment faster. For simplicity, we choose a fixed α with non-zero value.
- $\mathcal{R}$ is the reward, and it is a function of $\mathbb{E}[N_{PT}]$, $CV_{N_{PT}}$, ΔN_{PT}, and P_{ACB}. Due to the many possible combinations, we just show some of the possible state/reward combinations in Table 3.1. In general terms, we aim to maintain ACB off (i.e., $P_{\text{ACB}} = 1$) when there is low occupation, and to decrease the P_{ACB} value as traffic grows to reduce congestion. According to our observations, we consider that for $N_{UL} = 15$ uplink grants, a value of $\mathbb{E}[N_{PT}] > 10$ indicates congestion, and the rewards reflect this observation.
- γ is the discount factor that affects the presence of the sum of all future rewards in the current time slot. A very small γ implies that the relevance of future rewards in the algorithm is low compared with current ones.

An ϵ-greedy approach is used in selecting an action. Let ϵ be the exploration probability, $0 \leq \epsilon \leq 1$. Then, with probability ϵ, the algorithm chooses equal-probably an action from the remaining feasible actions (exploration). With probability $1 - \epsilon$, the algorithm will select the action with the highest $Q(s,a)$ value (exploitation). This is a trade-off between exploration and exploitation, where a higher ϵ will encourage more aggressive exploration for potentially better but yet-to-be-known action for a given state. In our experiments, the value of ϵ followed a linear function that went from 1 to 0. Note that Q-learning converges to the optimum action-values with probability 1 so

Table 3.1: Some Examples of Rewards Associated to Actions on RL-ACB

State			a'	$\mathcal{R}$
$\mathbb{E}[N_{PT}] \leq 3$,	$CV_{N_{PT}} < 0.4$,	$\Delta N_{PT} < 0$,	$P_{\text{ACB}}(s') = 1$	100
$\mathbb{E}[N_{PT}] \leq 3$,	$CV_{N_{PT}} < 0.4$,	$\Delta N_{PT} < 0$,	$P_{\text{ACB}}(s') \geq 0.7$	80
$\mathbb{E}[N_{PT}] < 7$,	$CV_{N_{PT}} < 0.4$,	$\Delta N_{PT} < 0$,	$P_{\text{ACB}}(s') \geq 0.5$	40
$\mathbb{E}[N_{PT}] < 7$,	$CV_{N_{PT}} < 0.4$,	$\Delta N_{PT} < 0$,	$P_{\text{ACB}}(s') \geq 0.3$	80
$\mathbb{E}[N_{PT}] \leq 10$,	$CV_{N_{PT}} \leq 0.2$,	$\Delta N_{PT} < 0$,	$P_{\text{ACB}}(s') \geq 0.05$	40
$\mathbb{E}[N_{PT}] > 10$,	$CV_{N_{PT}} \geq 0.2$,	$\Delta N_{PT} > 0$,	$P_{\text{ACB}}(s') = 1$	-100
$\mathbb{E}[N_{PT}] > 10$,	$CV_{N_{PT}} \geq 0.2$,	$\Delta N_{PT} > 0$,	$P_{\text{ACB}}(s') \geq 0.7$	-90
$\mathbb{E}[N_{PT}] > 10$,	$CV_{N_{PT}} \geq 0.2$,	$\Delta N_{PT} > 0$,	$P_{\text{ACB}}(s') \geq 0.5$	-60
$\mathbb{E}[N_{PT}] > 10$,	$CV_{N_{PT}} \geq 0.2$,	$\Delta N_{PT} > 0$,	$P_{\text{ACB}}(s') \geq 0.3$	-50
$\mathbb{E}[N_{PT}] < 7$,	$CV_{N_{PT}} \geq 0.4$,	$\Delta N_{PT} > 0$,	$P_{\text{ACB}}(s') \geq 0.05$	-20

long as all actions are repeatedly sampled in all states and the action-values are defined discretely [58, 59]. The RL-based ACB implementation is shown in Algorithm 1.

3.3.1 Performance evaluation

In this section, we evaluate the proficiency of our RL-based ACB scheme in terms of three KPIs, namely the probability to successfully complete the random access procedure, P_s; the number of preambles transmitted by the successfully accessed UEs (mean and percentiles), K; and the access delay (mean and percentiles), D.

A single cell environment is assumed to evaluate the network performance; the system accommodates both H2H and M2M UEs with different access request intensities. In order to assess the RL-based ACB scheme based on realistic H2H traffic behavior, we make use of call detail records (CDRs) obtained from a telco. The Italian operator Telecom Italia made available in 2014 a set of data from its network of the cities of Milan and Trento for what

Algorithm 1: RL-based ACB Scheme

Controller: Q-learning($\mathcal{S}, \mathcal{A}, \alpha, \mathcal{R}, \gamma, \epsilon$)

Input : $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of actions, α is the learning rate, $\mathcal{R}$ is the reward, γ is the discount factor, ϵ is the exploration probability

Local : real array $\mathbf{Q}[s, a]$, previous state s, previous action a

```
forall i = 1, 2, ... do
    if RAO(i) mod T_SIB2 = 0 then
        select action a' from A based on ε;
        observe reward R(s, a', s') and state s';
        update Q(s, a) by (3.1);
        update ac-BarringFactor in SIB2 with a'
    end
    s = s'
end
```

it defined as a *big data challenge* [60]. This data provides an intensity measure of data traffic for a constrained area, aggregated in periods of 10 minutes during two months (November and December of 2013). This data is very useful to evaluate the temporal and geographical distributions of H2H traffic for a specific service (data, voice, SMS).

According to [61], the impact of data traffic on the RACH procedure can be 50 times higher than that of voice traffic, due mainly to the short-timed, high-frequency, low-data volume connections of apps in background mode. Therefore, it is necessary to pre-process this data. Also in [61], it is stated that a base station (eNodeB) can support up to 55 EUTRAN radio access bearer setups per second in high load scenarios. Hence, we use this value as a reference, and normalize the original data accordingly. Since data from H2H traffic is aggregated every 10 minutes, we assume that during this period the traffic is constant. Considering H2H traffic as background traffic, we add M2M traffic in each period and evaluate a heavy-loaded scenario (30 000

M2M UEs). This M2M traffic follows a Beta(3,4) distribution over 10 seconds (2000 RAOs) as described in [5]. We measure the KPIs once the M2M UEs have completed their random access procedure.

We consider the typical PRACH configuration, *prach-ConfigIndex 6*, in conformance to the LTE-A specification for these kind of studies [5, 23], where the subframe length is 1 ms and the periodicity of RAOs is 5 ms. Also $R = 54$ out of 64 available preambles are used for the contention-based random access and the maximum number of preamble transmissions per UE, *preambleTransMax*, is set to 10. Table 2.3 lists additional parameters used throughout our analysis (unless otherwise stated). Although there is a high variation of traffic in H2H communications according to the day, time, or specific geographical position of the cell, its intensity is significantly smaller than that of M2M traffic. Hence, we focus on one of the most occupied cells found in the traces (cell 5161) located in the center of the city, near the Milan Cathedral at 4:20 pm, which is the time with the highest utilization on November 16.

Fig. 3.2 depicts the temporal distribution of access requests per RAO on the above mentioned cell with a burst of M2M traffic. As can be seen, a congestion control mechanism is necessary; besides, such a high number of preamble transmissions is the consequence of the fact that the higher the number of preamble transmissions in a RAO, the lower the probability of a successful preamble transmission. This, in turn, increases the probability of preamble re-transmissions in the following RAOs, hence the probability of a successful preamble transmission is further reduced.

Fig. 3.3 shows the access request per RAO when the static ACB with parameters $P_{\text{ACB}} = 0.5$ and $T_{\text{ACB}} = 4\,\text{s}$ is implemented. These parameter values were picked based on the previous analysis conducted in Section 2.6 where it was identified that the combination of low values of T_{ACB} with high values of P_{ACB} leads to a reduction in the access delay; particularly, the lowest access delay for a highly congested scenario given a requirement of $P_s \geq 0.95$ is achieved when $P_{\text{ACB}} = 0.5$ and $T_{\text{ACB}} = 4\,\text{s}$. However, the number of collisions is still high because the average number of preamble transmissions

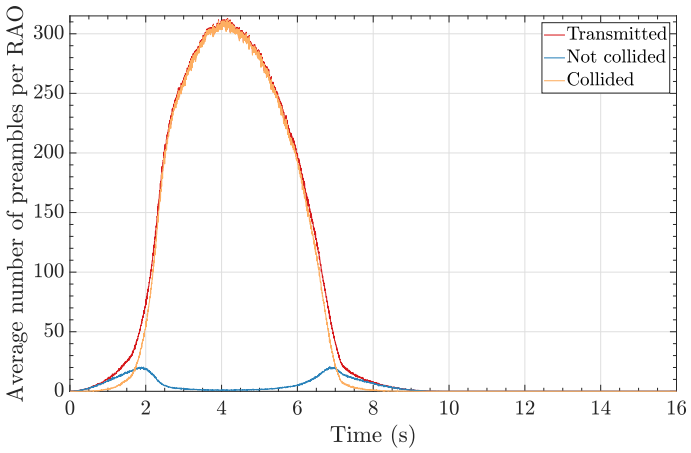


Figure 3.2: Average number of preambles per RAO when H2H UEs and massive M2M UEs attempt to access. No access control implemented.

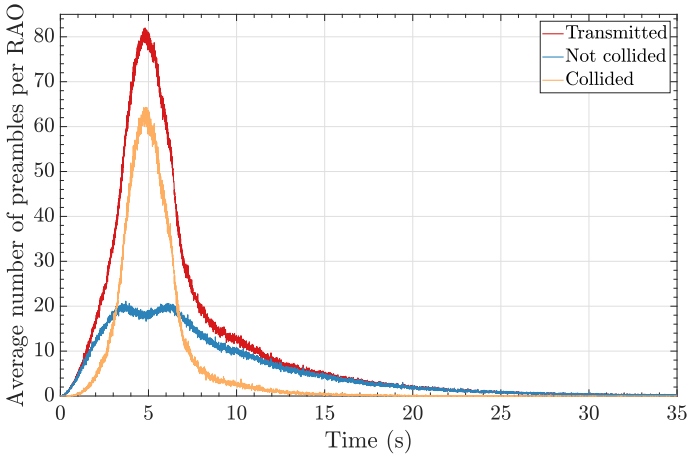


Figure 3.3: Average number of preambles per RAO when H2H UEs and massive M2M UEs attempt to access, static ACB(0.5,4 s) is implemented.

surpasses the RACH capacity which is 20.05 in a scenario with 54 available preambles like this one [2, 49].

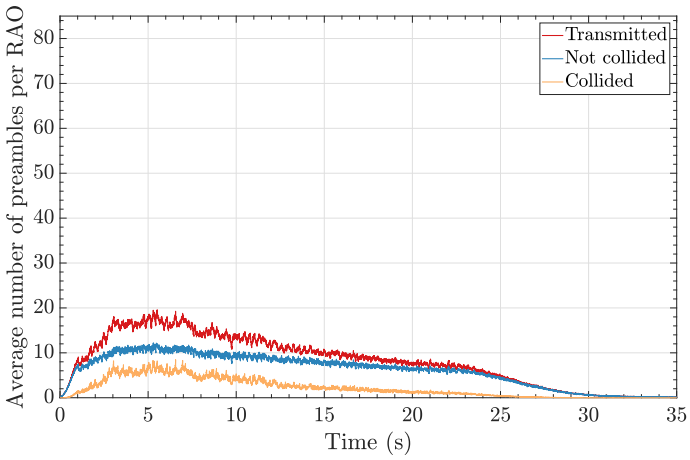


Figure 3.4: Average number of preambles per RAO when H2H UEs and massive M2M UEs attempt to access, RL-based ACB is implemented.

For the experiments associated with Q-learning, the algorithm was trained for one day (November 15); the parameter values used for this period were $\alpha = 0.15$, $\gamma = 0.7$, and a linear function from 1 to 0 for ϵ. We consider this training period significant since it represents around 6×10^5 epochs. Then, we tested the algorithm on November 16 on the cell with the highest occupation; we used different seeds for the M2M access distribution, which allowed us to test 200 different experiments. The results shown in Fig. 3.4 represent the mean of these 200 experiments. As can be seen, the number of collisions was greatly reduced and it is consistently smaller than the number of successful transmissions. This is due to the fact that in our rewards system there was a strong bias towards avoiding congestion. As a result, the number of successful accesses and the number of first preamble transmissions are very close for the whole measured period. Also, the total number of preamble transmissions was considerably reduced when compared to the LTE-A system without access control, and to the LTE-A system with static ACB. More importantly, this reduction was achieved under dynamic conditions and by adapting P_{ACB} accordingly.

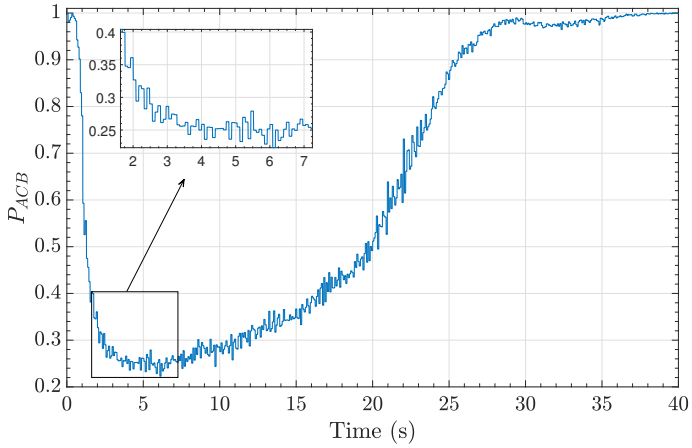


Figure 3.5: Evolution of P_{ACB} as a function of time using Q-learning.

Fig. 3.5 showcases the mean value of P_{ACB} as it adapts to different rates of UE arrivals. It can be seen that in the first RAO, P_{ACB} is equal to 1; then, it quickly decreases to around 0.25 when the number of total preamble transmissions rises, but then grows again as the traffic diminishes, until it goes back to 1, where it settles. It should be noted that P_{ACB} changes dynamically with a granularity of T_{SIB2}, that is 16 RAOs. Hence, through an appropriate setting of the Q-learning parameters, it is possible to reduce collisions, although the cost is a higher delay.

In Table 3.2, we can see different statistics for the same cell, during the same time period, for the two different access control schemes. We separate the results for each type of service (M2M and H2H), and obtain the KPIs of the network defined at the begining of this section.

It is evident from the results that the network without ACB suffers in terms of P_s and K. However, it has the smallest delay. On the other hand, our proposed ACB scheme based on Q-learning reaches the best P_s, with practically a 100% success. This is consistent with the results seen earlier on Fig. 3.4 and shows an improvement over the solution with static ACB. Also, the

Table 3.2: KPIs Obtained for LTE-A and Different ACB Implementations. Massive M2M + H2H Traffic

Key Performance Indicator		No ACB		ACB(0.5, 4 s)		RL-based ACB	
		M2M	H2H	M2M	H2H	M2M	H2H
Success probability	P_s	0.31	0.60	0.97	0.99	0.99	1
Number of preamble transmissions, K	$\mathbb{E}[K]$	3.46	2.36	2.49	1.57	1.85	1.62
	K_{95}	8.58	6.71	6.31	2.63	6.17	2.71
	K_{50}	1.99	1.19	1.42	1.07	1.00	1.00
	K_{10}	1.00	1.00	1.00	1.00	1.00	1.00
Access delay, D [ms]	$\mathbb{E}[D]$	67.94	45.01	4 162.9	3 512.5	7 657.6	3 463.9
	D_{95}	182.70	144.63	15 839.0	13 650.0	19 924.0	15 164.0
	D_{50}	47.10	30.14	2 955.0	59.90	6 534.0	45.00
	D_{10}	17.85	16.80	21.30	16.80	17.98	16.80

Q-learning solution reduces the mean number of preambles transmitted for M2M communications, which are the ones responsible for the bursty traffic. Furthermore, our solution is able to reduce this KPI without considerably increasing the mean number of preamble transmissions for H2H traffic.

This is important since one of the main requirements when introducing M2M communications into an LTE-A network is that it does not affect the preexisting H2H UEs. In fact, the mean access delay for H2H users is lower for the Q-learning scheme than in the solution with static ACB. However, as expected, there is a trade-off, and this is reflected on an increment on the delay for M2M communications. This is expected since as it was shown in Fig. 3.4, the collisions were considerably reduced.

3.4 Estimating the number of UEs in backoff state approach

The ultimate goal of the ACB is to control the arrival of M2M UEs to the RACH so that the number of transmitted preambles stays below a certain threshold. To achieve this, the eNodeB must tune in real-time the P_{ACB} value accordingly to the traffic conditions. Fig. 3.6 illustrates the states when performing the ACB mechanism [4, 21] jointly with the random access procedure. We propose to implement the P_{ACB} updating based on an estimate of the number of UEs that are in backoff state. In Section 3.4.1 we derive analytically the estimation of the number of UEs in backoff state and in Section 3.4.2 we describe the mechanism for tuning the P_{ACB} dynamically.

3.4.1 Estimation of the number of UEs in backoff state

At the RAO (i), let $n_b(i)$ be the estimate number of UEs in backoff state, $n_s(i)$ the number of UEs that accessed successfully, $n_t(i)$ the number of UEs that transmitted a preamble and $n_r(i)$ the number of UEs in backoff state that retransmitted. The estimated number of UEs in backoff state at RAO (i) (see Fig. 3.7) is given by

$$n_b(i) = n_b(i-1) - n_r(i) + n_t(i) - n_s(i). \tag{3.2}$$

UEs that fail the access, mainly because they reach the *preambleTransMax* value, are not considered in Eq. (3.2). Therefore, the estimation presented here is valid only when the success probability is high enough, which is a reasonable assumption when congestion control performs effectively, as we show in the following sections.

The eNodeB knows $n_s(i)$ by observing the preambles that have been detected and decoded correctly. Since the number of uplink grants that the eNodeB can send in an RAO is limited to N_{UL} grants, $n_s(i)$ is given by

$$n_s(i) = \min\{\text{preambles decoded}, N_{UL}\}. \tag{3.3}$$

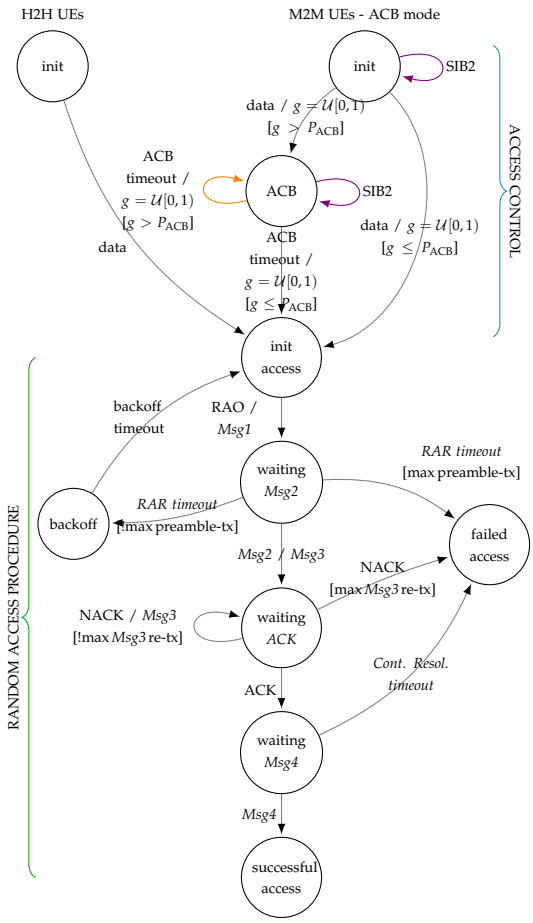


Figure 3.6: State transition diagram of the random access procedure when access control is implemented for M2M UEs.

The value of $n_t(i)$ is unknown because, in our collision model, the preambles that collided (transmitted by more than one UE) are not decoded, and therefore the eNodeB does not know how many UEs transmitted them. Instead, we will use the estimated expectation of n_t, obtained from the number of used preambles n_u (preambles transmitted at least by one UE). Let

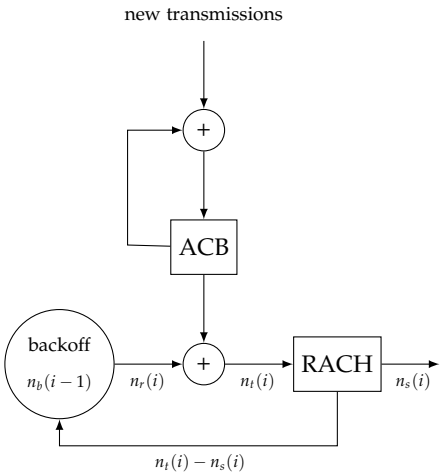


Figure 3.7: Variables and traffic flows used in the estimation of $n_b(i)$.

$Y_j \in \{0, 1\}$ be the random variable that denotes the transmission of preamble j given that the total number of transmissions is n_t. Then, $Y_j = 0$ when preamble j has not been transmitted by any UE, and $Y_j = 1$ otherwise. The probabilities of Y_j are

$$\begin{cases} \mathbb{P}\{Y_j = 0\} = \left(1 - \frac{1}{R}\right)^{n_t}, \\ \mathbb{P}\{Y_j = 1\} = 1 - \left(1 - \frac{1}{R}\right)^{n_t}, \end{cases} \tag{3.4}$$

where R is the number of available preambles in the system. The expected value of Y_j is given by

$$\mathbb{E}\{Y_j\} = 0 \cdot \mathbb{P}\{Y_j = 0\} + 1 \cdot \mathbb{P}\{Y_j = 1\} = 1 - \left(1 - \frac{1}{R}\right)^{n_t}. \tag{3.5}$$

Then, the number of used preambles, n_u, is

$$n_u = \sum_{j=0}^{R} Y_j, \tag{3.6}$$

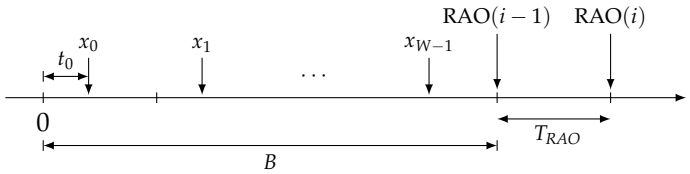


Figure 3.8: Backoff starting times.

and its expected value is

$$\mathbb{E}\{n_u\} = \mathbb{E}\left\{\sum_{j=0}^{R} Y_j\right\} = R\left[1 - \left(1 - \frac{1}{R}\right)^{n_t}\right]. \tag{3.7}$$

If we assume that $\mathbb{E}\{n_u\}$ changes slowly, it can be estimated from a short term time average of n_u. Let $\hat{n}_u(i)$ be an estimate of the time average of n_u obtained by exponential smoothing of n_u at RAO (i). Then,

$$\mathbb{E}\{n_u\}(i) \approx \hat{n}_u(i) = \alpha\, \hat{n}_u(i-1) + (1-\alpha)\, \hat{n}_u(i) \tag{3.8}$$

with $\alpha < 1$, and from Eq. (3.7), $n_t(i)$ is

$$n_t(i) = \frac{\log\left(1 - \frac{\mathbb{E}\{n_u\}(i)}{R}\right)}{\log\left(1 - \frac{1}{R}\right)}. \tag{3.9}$$

The value of $n_r(i)$ is estimated from $n_b(i-1)$. Firstly, from $n_b(i-1)$ we estimate the expected number of UEs that entered the backoff state during the period of length B (maximum backoff duration) previous to RAO$(i-1)$. Let n_A denote this number. Secondly, from n_A we estimate the expected number of UEs that entered the backoff state during the period of length $B + T_{RAO}$ previous to RAO (i), that we denote as n_B. Finally, from n_B we estimate the expected value of $n_r(i)$.

Let X be the random variable that represents the UE backoff start time in the interval $[0, B]$. The UEs enter the backoff state at a fixed time t_0 after the transmission of a preamble, where t_0 depends on the RACH configuration

parameters (see Table 2.3). Besides, preambles are transmitted at a RAO subframe, so all values of X are grouped in subframe intervals every T_{RAO}. We approximate these values to the discrete set $\{x_0 \dots x_{W-1}\}$, being $x_i = t_0 + i\ T_{RAO}$, as shown in Fig. 3.8. We assume that at any time window $B = W\,T_{RAO}$, all backoff times have the same probability (i.e., $\mathbb{P}(X = x_i) = 1/W$).

Let T_{BO} be the random variable that represents the backoff duration, with uniform distribution given by Eq. (2.1). If we denote by p the probability that a UE starting backoff in the interval $[0, B]$ is still in backoff at time B, p is computed as

$$\begin{aligned} p &= \sum_{w=0}^{W-1} \mathbb{P}(X = x_w, x_w + T_{BO} > B) = \frac{1}{W} \sum_{w=0}^{W-1} \frac{x_w}{B} \\ &= \frac{1}{W} \sum_{w=0}^{W-1} \frac{w\,T_{RAO} + t_0}{B} = \frac{1}{W}\frac{1}{B} W \frac{x_0 + x_{W-1}}{2} \\ &= \frac{1}{2B}\left(t_0 + (W-1)T_{RAO} + t_0\right) \\ &= \frac{1}{2}\left(1 - \frac{T_{RAO}}{B}\right) + \frac{t_0}{B}. \end{aligned} \tag{3.10}$$

The number of UEs in backoff at RAO $(i-1)$ is a random variable that follows a binomial distribution with success probability p, n_A attempts, and expected value $p\,n_A$. Note that we do not consider those UEs that ended the backoff and started a new one before RAO$(i-1)$. But, if the congestion control performs effectively, the probability of a second backoff will be low, and the error that is incurred by this approximation is negligible. If we take $n_b(i-1)$ as the expected value of the number of UEs in backoff, we can estimate n_A from

$$n_b(i-1) = p\,n_A. \tag{3.11}$$

We also assume that users will start backoff at $x_{W-1} + T_{RAO}$ with the same probability as in $x_i, i \in \{0, \dots, W-1\}$. Then, n_B can be estimated from

$$n_B = n_A + \frac{n_A}{W} = \left(1 + \frac{T_{RAO}}{B}\right) n_A. \tag{3.12}$$

Let Z be the discrete random variable that represents the UEs backoff start time during the interval $[0, B + T_{RAO}]$, with values $\{z_0 \dots z_W\}$ and probabilities $\mathbb{P}(Z = z_i) = 1/(W+1)$. If we denote by q the probability that a UE starting backoff in the interval $[0, B + T_{RAO}]$ retransmits in the interval $[B, B + T_{RAO}]$, q is computed as

$$\begin{aligned} q &= \sum_{w=0}^{W} \mathbb{P}(Z = z_w, B < z_w + T_{BO} < B + T_{RAO}) \\ &= \frac{1}{W+1}\mathbb{P}(T_{BO} > B - z_0) \\ &\quad + \frac{1}{W+1}\sum_{w=1}^{W-1} \mathbb{P}(B - z_w < T_{BO} < B - z_w + T_{RAO}) \\ &\quad + \frac{1}{W+1}\mathbb{P}(T_{BO} < B - z_W + T_{RAO}) \\ &= \frac{1}{W+1}\left(\frac{t_0}{B} + \sum_{w=1}^{W-1}\frac{T_{RAO}}{B} + \frac{B + T_{RAO} - (WT_{RAO} + t_0)}{B}\right) \\ &= \frac{1}{W+1}\left((W-1)\frac{T_{RAO}}{B} + \frac{T_{RAO}}{B}\right) = \frac{T_{RAO}}{B + T_{RAO}}. \end{aligned} \tag{3.13}$$

The number of UEs in backoff that retransmit at RAO (i) is a random variable that follows a binomial distribution with success probability q, n_B attempts, and expected value $q\, n_B$. If we take $n_b(i)$ as the expected value of this variable, we can estimate $n_r(i)$ as

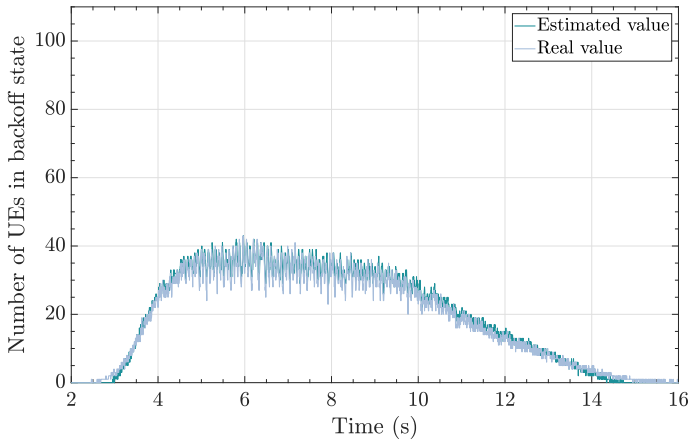
$$n_r(i) = q\, n_B, \tag{3.14}$$

and, from Eqs. (3.10)–(3.14), it follows that

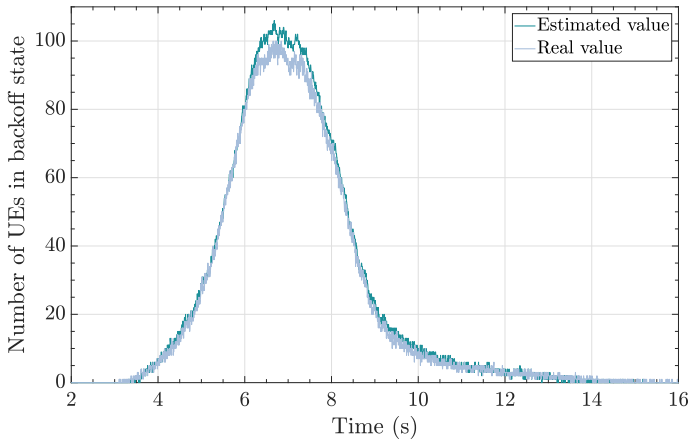
$$n_r(i) = \frac{2\ T_{RAO}}{B - T_{RAO} + 2\ t_0}\, n_b(i-1). \tag{3.15}$$

Finally, $n_b(i)$ is computed by replacing Eqs. (3.15), (3.9), and (3.3) in Eq. (3.2).

Fig. 3.9 shows an example, obtained by discrete-event simulation, of the performance of this estimator. In this example, the estimator operates during



(a)



(b)

Figure 3.9: Performance of the estimator during a congestion episode. (a) Dynamic ACB. (b) Static ACB.

a congestion episode induced by the M2M traffic benchmark described in Section 3.4.3, with the UEs implementing ACB. As can be seen, the error in the estimation is minimal.

3.4.2 Dynamic barring rate tuning

In this section, we describe a mechanism for tuning the barring rate dynamically using the information provided by the estimation of the number of users in backoff state, $n_b(i)$. We assume that the eNodeB has no information about the number of new arrivals. Nevertheless, by estimating $n_b(i)$, the expected number of retransmissions, $n_r(i)$, is also estimated. From this, the unused capacity of the RACH can be evaluated and the incoming traffic can be modulated accordingly, that is, by computing P_{ACB} as a function of $n_r(i)$: $P_{\text{ACB}} = f(n_r(i))$.

The number of access requests per RAO that maximizes the expected number of UEs that access the channel successfully is given by [2]

$$n_t^*(R) = [\log(R/(R-1)]^{-1}. \tag{3.16}$$

If $n_r(i) \geq n_t^*(R)$, all new arrivals should be barred, so $f(n_r(i) \geq n_t^*(R)) = 0$. On the other hand, if $n_r(i) = 0$ there is no reason to reduce the incoming traffic, so $f(0) = 1$. Between these two boundary conditions, $f(\cdot)$ should be a decreasing function. We have tested by simulation linear, quadratic, and cubic functions and we have found that, although all of them work quite well, the cubic function shown in Fig. 3.10 provides a better performance. Thus, we propose to compute P_{ACB} as

$$P_{\text{ACB}}(i) = f(n_r(i)) = \frac{2\, n_r(i)^3}{n_t^*(R)^3} - \frac{3\, n_r(i)^2}{n_t^*(R)^2} + 1. \tag{3.17}$$

The value of P_{ACB} is notified to the UEs by the eNodeB through the SIB2 message with a periodicity of $T_{SIB2} = 80\,\text{ms} \equiv 16\,\text{RAOs}$.

Finally, the dynamic ACB implementation can be summarized as

- Every RAO (i),
  1. Estimate $\mathbb{E}\{n_u\}(i)$ with Eq. (3.8),
  2. Estimate $n_t(i)$ with Eq. (3.9),

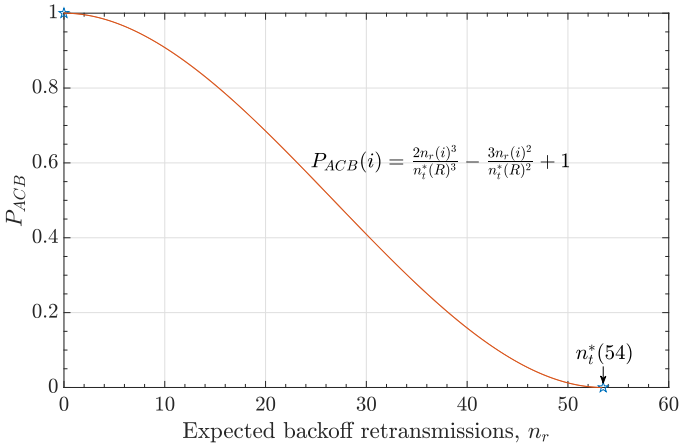


Figure 3.10: Computation of P_{ACB} from the expected number of retransmissions.

3. Estimate $n_r(i)$ with Eq. (3.15),
4. Estimate $n_b(i)$ with Eq. (3.3).

- Every T_{SIB2},
  1. Compute $P_{\text{ACB}}(i)$ with Eq. (3.17),
  2. Update $P_{\text{ACB}}(i)$ in SIB2.

3.4.3 Performance evaluation

In this section, we evaluate the performance of the proposed dynamic ACB scheme with $T_{\text{ACB}} = 4\,\text{s}$, and compare it with the performance of a static ACB with optimal parameters ($P_{\text{ACB}} = 0.5$ and $T_{\text{ACB}} = 4\,\text{s}$). The parameters for static ACB are chosen based on the previous analysis conducted in Section 2.6 where we showed that the lowest access delay for a highly congested scenario given a requirement of $P_s \geq 0.95$ is achieved when $P_{\text{ACB}} = 0.5$ and $T_{\text{ACB}} = 4\,\text{s}$. We measure three KPIs, namely the probability to successfully

complete the random access procedure, P_s; the mean number of preamble transmissions needed by the UEs to successfully complete the random access procedure, K; and the access delay (mean and percentiles), D. These KPIs are in conformance with the 3GPP directives [5] to assess the efficiency of the LTE-A random access procedure with M2M communications.

A single cell environment is assumed to evaluate the network performance. The system conveys traffic from both H2H and M2M UEs. The access requests of H2H UEs are distributed uniformly over time with a mean arrival rate of $\lambda_H = 1\,\text{arrivals/s}$. The M2M requests follow a $Beta(3,4)$ distribution over a period of 10 s, according to the traffic model 2 specified by the 3GPP in [5]. This traffic model can be seen as an extreme scenario in which a vast number of M2M UE arrivals occur in a highly synchronized manner (e.g., after an alarm that activates them).

We developed a discrete-event simulator that fully reproduces the behavior of UEs, eNodeB, and RACH during the random access procedure described in Section 2.3.1. We consider the typical PRACH configuration for these kind of studies, *prach-ConfigIndex 6*, in conformance to the LTE-A specification [5, 23], with the parameter values listed in Table 2.3. The value of each KPIs is obtained as the mean of 100 simulation runs. Each simulation run uses a different random seed and ends when all the M2M UEs have completed their random access procedure.

Fig. 3.11 depicts the temporal distribution of preamble transmissions during a congestion episode of $N_M = 30\,000$ M2M UEs and no congestion control. As can be seen, an access control mechanism is required for alleviating congestion. Such a high number of preamble transmissions is the consequence of the fact that the higher the number of preamble transmissions in a RAO, the lower the probability of a successful preamble transmission. This, in turn, increases the probability of preamble retransmissions and reduces dramatically the probability of successful access.

Fig. 3.12 shows the above scenario for the static ACB with $P_{\text{ACB}} = 0.5$ and $T_{\text{ACB}} = 4\,\text{s}$. Now, the total number of preamble transmissions is reduced

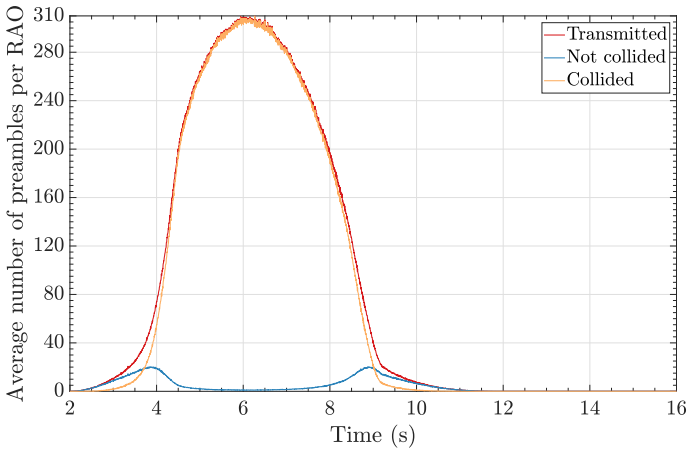


Figure 3.11: Average number of preambles per RAO when H2H UEs and massive M2M UEs ($N_M = 30\,000$) attempt to access. No access control.

while the number of not collided transmissions increases, what results in a higher probability of successful access. But the number of collisions is still high because it fails to maintain the preamble transmission rate close to the optimal number of access requests per RAO that maximizes the expected number of successful access, $n_t^*(R)$, which is approximately 53 with $R = 54$ available preambles.

Fig. 3.13 illustrates the performance of our proposed dynamic ACB scheme in the same scenario. We see that the total number of preamble transmissions and the number of collided transmissions are both considerably reduced when compared to the static ACB. More importantly, this reduction is achieved under dynamic conditions and by adapting P_{ACB} accordingly. Note that most of the time the preamble transmissions rate is maintained close to the optimal value ($n_t^*(R) \approx 53$) and the not collided preambles rate is maintained close to the RACH capacity (approx. 20 [49, Eq. (4)]), using the available resources efficiently.

Fig. 3.14 shows how P_{ACB} is tuned by the eNodeB with dynamic ACB op-

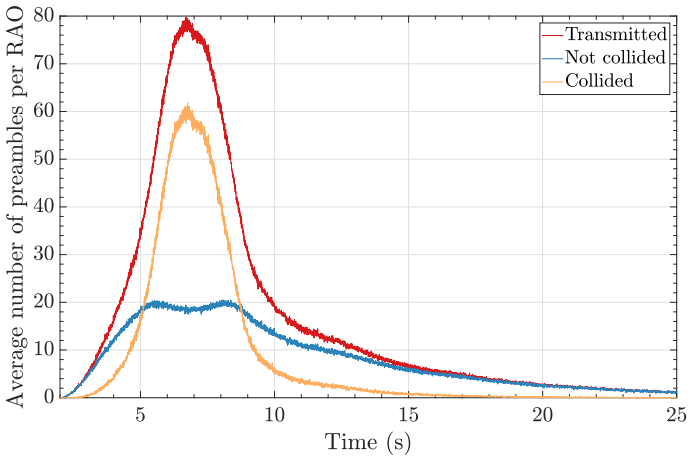


Figure 3.12: Average number of preambles per RAO when H2H UEs and massive M2M UEs ($N_M = 30\,000$) attempt to access. Static ACB (0.5, 4 s).

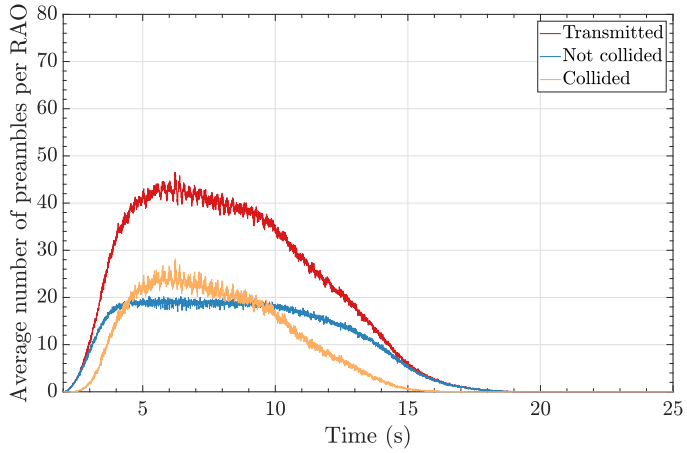


Figure 3.13: Average number of preambles per RAO when H2H UEs and massive M2M UEs ($N_M = 30\,000$) attempt to access. Proposed dynamic ACB.

erating during the congestion episode. As can be seen, before the congestion starts, $P_{\text{ACB}} = 1$; it gradually decreases when the number of preamble trans-

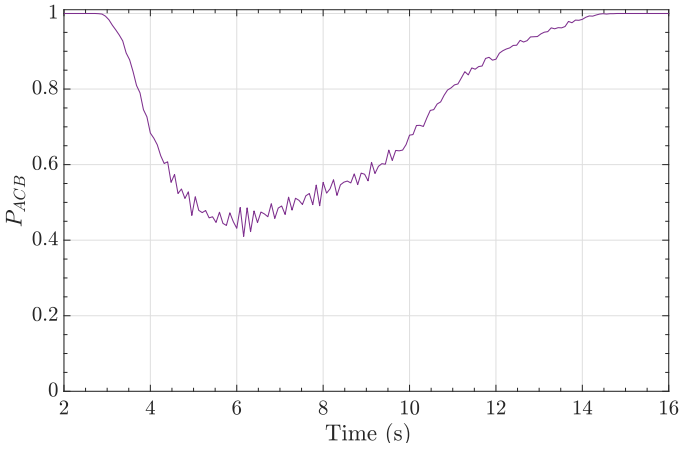


Figure 3.14: Evolution of P_{ACB} in dynamic ACB during a congestion episode.

missions rises, and grows again as the traffic diminishes, until it goes back to 1. Note that P_{ACB} changes dynamically with a granularity of T_{SIB2}.

Fig. 3.15 depicts the successful access probability, P_s, for both ACB types with different M2M traffic intensities (N_M). As can be seen, in heavy-loaded scenarios ($N_M > 26\,000$) the dynamic ACB provides a $P_s \geq 0.99$, clearly outperforming the static one. In terms of mean number of preamble transmissions needed to successfully complete the random access procedure, Fig. 3.16 shows that the dynamic ACB slightly increases this metric when ($30\,000 > N_M$).

In Fig. 3.17, we illustrate the mean access delay of both ACB types. We can observe that the dynamic ACB decreases notably the mean access delay. This reduction is even greater in low-loaded scenarios, since our solution uses efficiently the available RACH resources. Therefore, for a given QoS in terms of access success probability, the dynamic ACB can handle congestion episodes offering shorter access delay with negligible impact on the number of preamble transmissions.

In Table 3.3, we can see several statistics for the three schemes: no control,

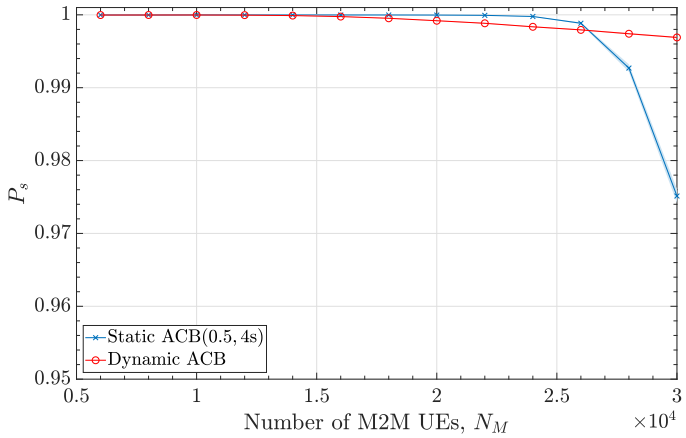


Figure 3.15: Successful access probability of the LTE-A network for static ACB and the proposed dynamic ACB.

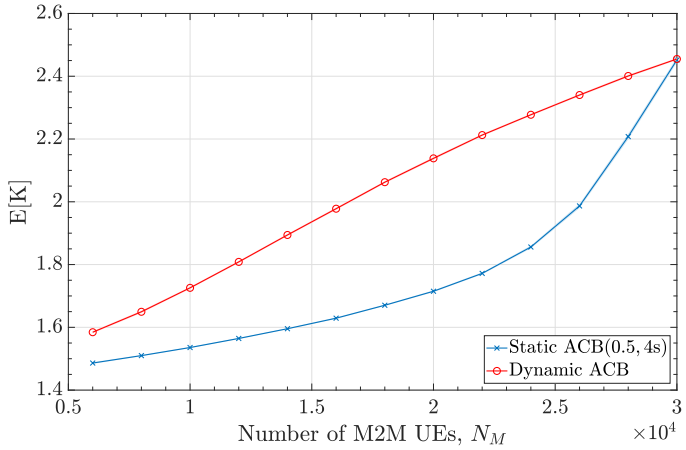


Figure 3.16: Expected number of preamble transmissions per RAO of the LTE-A network for static ACB and the proposed dynamic ACB.

static ACB, and dynamic ACB. For each of them, results for a heavy-loaded traffic scenario ($N_M = 30\,000$) are shown. We show the KPIs of the network

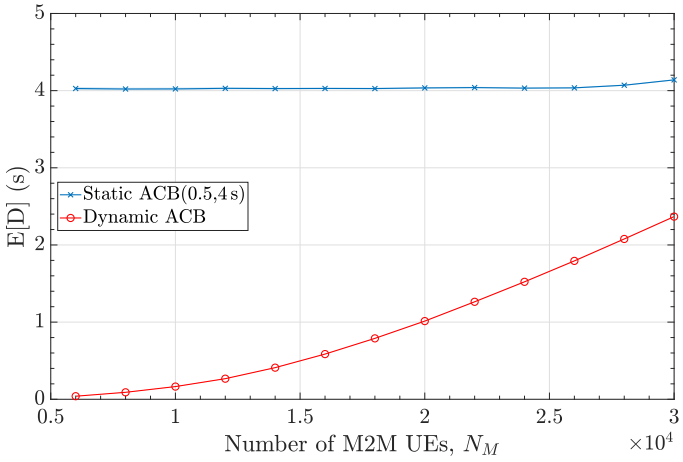


Figure 3.17: Expected access delay of the LTE-A network for static ACB and the proposed dynamic ACB.

both H2H and M2M UEs. It is evident from the results that the network with no access control suffers in terms of P_s and K. However, the access delay is moderate. On the other hand, our proposed dynamic ACB scheme shows an improvement over the solution with static ACB, providing a $P_s \geq 0.99$ for both H2H and M2M communications while reducing considerably the M2M access delays without increasing the H2H access delays. This is important since one of the main requirements when introducing M2M communications into an LTE-A network is that it does not affect the preexisting H2H UEs.

3.5 Highlights

We have proposed two solutions that allow efficient congestion control, facilitating the coexistence of H2H and M2M traffic. Both solutions conform with current system specifications.

First, we have proposed a dynamic mechanism based on reinforcement learning for tuning the barring rate. We considered a hybrid scenario with

Table 3.3: KPIs Obtained for LTE-A and Different ACB Implementations in a Heavy-Loaded Traffic Scenario.

Key Performance Indicator		No ACB		ACB(0.5, 4 s)		Dynamic ACB	
		H2H	M2M	H2H	M2M	H2H	M2M
Success probability	P_s	0.66	0.31	0.99	0.97	0.99	0.99
Number of preamble transmissions, K	$\mathbb{E}[K]$	2.15	3.45	1.60	2.45	2.03	2.45
Access delay, D [ms]	$\mathbb{E}[D]$	46	71	32	4 111	42	2 367
	D_{95}	141	186	67	15 744	104	8 242
	D_{80}	59	117	43	7 865	56	4 742
	D_{10}	17	18	16	22	17	21

both M2M and H2H communications. To provide a more realistic analysis of this type of scenarios, we used CDRs to model the H2H traffic; the M2M traffic is modeled according to the LTE-A specifications. Besides adapting the barring rate to sudden changes in traffic intensity, the proposed solution adjusts this traffic to the random access channel capacity consequently reducing the number of collisions and enhancing the probability of successful access, P_s. Also, our results show that although the enhancement of P_s can increase the access delay, it does not have a significant impact on H2H traffic, which is a necessary condition for the implementation of massive M2M communications. The Q-learning algorithm is aimed at reducing collisions, and therefore it has a slight impact on access delay.

Second, we have proposed a dynamic ACB mechanism that adjusts the barring rate in real-time, based on the estimation of the number of UEs that are in backoff sate. We have evaluated this solution by means of extensive discrete-event simulations, and compared its performance with that of a static ACB with optimal configuration. We conclude that, in a heavy-loaded scenario, the proposed dynamic ACB increases noticeably the probability of successful access while shorting the access delay, and number of preamble transmissions. Moreover, we have checked that the proposed dynamic ACB

does not have a significant impact on H2H traffic performance.

Comparing both proposed solutions in a heavy-loaded traffic scenario, the former offers a $P_s \geq 99\%$ at the cost of longer access delay whereas the latter offers a $P_s \geq 99\%$ with shorter access delay but increasing slightly the number of preamble transmissions.

Chapter 4

SDN-based Architecture for Reliable IoT Connectivity within 5G Systems

4.1 Introduction

Internet-of-Things (IoT) is a ubiquitous network of interconnected objects that harvest information from diverse environments, interact with the physical world, and use existing Internet infrastructure to provide services for information transfer, analytics, and applications. IoT services have enabled a wide range of emerging applications, such as environment monitoring, smart grid, smart city, smart transportation, e-health, and smart home. However, current IoT solutions are inherently hardware-based and rely on closed and inflexible architectural design. This fact imposes significant challenges on adopting new communication and networking technologies, also it prevents the provision of truly differentiated services to highly diverse IoT applications, inducing great barriers to integrating heterogeneous IoT devices.

In this chapter, to adequately address the above challenges in 5G IoT, we introduce a new architecture proposed for wireless software-defined networks, the so-called SoftAir [3] which decouples control and data planes for open, programmable, and virtualizable wireless forwarding infrastructure.

The data plane consists of software-defined RANs (SD-RANs) and software-defined core networks; the control plane has network management tools and user applications. Then, we propose software-defined gateways (SD-GWs) as local IoT controllers in SoftAir. They can be deployed for satisfying the massive connectivity and diverse traffic generated by a vast number of IoT devices. SD-GW, serving as the bridge between IoT networks and SD-RANs, provide intensive data aggregation from heterogeneous IoT devices, manage and orchestrate IoT communication, and perform protocol conversions between IoT networks and SD-RANs. Our solution overcomes the limitations of existing commercial wireless networks by offering five core properties: programmability, cooperativeness, virtualizability, openness, and visibility. These five properties provide functionalities that are essential to enable 5G wireless communication networks and support emerging IoT applications and services. We consider a likely IoT scenario based on several wireless sensor networks (WSNs) that provide IoT services through the SoftAir system.

The main contributions of this chapter are summarized as follows.

- We present the SoftAir architecture to provide IoTs connectivity that exploits the emerging features in wireless communications. A study that explores the interactions of communication functionalities for an IoT scenario based on WSNs is provided.
- We design a heterogeneous cross-layer solution for the SD-GW aiming to fulfill a predefined level of quality of service (QoS), efficient energy consumption, high system performance, and reliable connectivity.
- We develop an optimization framework that achieves optimal energy savings and throughput gain concurrently in WSNs while maximizing the SD-GW rate coverage with mmWave remote radio heads (RRHs) coordination in SoftAir.

The rest of the chapter is organized as follows. Section 4.2 presents the motivation and related work in this context. Section 4.3 describes the SoftAir architecture for IoT communications. Section 4.4 presents the heterogeneous

cross-layer optimization design in SD-GWs that integrates IoTs and SD-RAN of SoftAir. Section 4.5 gives the performance evaluation, and Section 4.6 concludes the chapter.

4.2 Motivation and related work

With the explosion of devices connected through IoT, traditional network architectures will not be able to handle both the number of devices and the volume of data they will be draining into the network. Moreover, current IoT solutions rely on low-power wide area (LPWA) networks [62], which complement traditional cellular and short-range wireless technologies in addressing IoT applications. Several technologies, such as Long Range (LoRa), narrow band IoT (NB-IoT), SIGFOX, have been developed and designed solely for applications with very limited demands on throughput, reliability, or QoS [63]. However, without a central regulation among these LPWA technologies, existing IoT solutions cannot support highly diverse QoS requirements from increasing 5G IoT applications. Due to currently fixed and hardware-based infrastructure, no existing work has considered the joint architectural design of IoT networks and SD-RANs, and the provision of reliable and efficient upstream/downstream IoT transmissions. Another challenge is to efficiently manage the load of traffic and the network resources in the 5G era, to avoid a possible collapse of the network, and to allow the coexistence of different services with different QoS requirements in a scalable and efficient manner.

This motivates us to develop a cross layer framework and propose a joint optimization of protocols crossing different layers from the IoTs to the SD-RAN according to the devices' QoS requirements and system constraints. Thus, we provide a solution for various performance requirements of applications to handle the heterogeneity of IoT devices. Specifically, inspired by wireless software-defined networking [9, 64], we first propose the SoftAir architecture to support a unified software-defined platform for QoS-aware IoT systems and SD-RANs with millimeter-wave transmissions. Next, the SD-

GWs are designed in SoftAir to explore the interactions between two-types of networks (i.e., IoTs and SD-RANs) and enable cross-layer solutions that simultaneously achieve optimal energy savings and throughput gain in IoTs and maximum sum-rates in SD-RANs. The proposed SD-GW in the SoftAir architecture will manage the sporadic communications from a myriad of the heterogeneous IoT devices and provide local offloading. Furthermore, with the introduction of IPv6, the vast increase in the number of connected devices is properly addressed and the SD-GW can be used to send IoT data to other devices connected to the Internet.

4.3 SoftAir architecture for 5G IoT

SoftAir [3] is a unified software-defined platform for 5G systems with network management tools and customized applications of service providers or virtual network operators. It would enable IoT applications to access the data and control the devices without the knowledge of the underlying infrastructure. SoftAir follows a distributed RAN architecture composed of three main parts: (i) the centralized base band server (BBS) pool, which connects to the core network via backhaul links and consists of software-defined base stations (SD-BSs) from real-time virtualization technology for software-implemented baseband units (e.g., digital processing tasks); (ii) RRHs plus antennas, which are remotely controlled by SD-BSs and serve SD-GWs' transmissions; and (iii) low-latency high-bandwidth fronthaul links (fiber or microwave) using the common public radio interface (CPRI) for an accurate, high-resolution synchronization among RRHs.

Extended from our preliminary study in [3], Fig. 4.1 depicts an example of the SoftAir-based architecture for 5G IoTs. It consists of three domains: sensing, network, and application. The sensing domain enables *things* to interact and communicate with themselves and with the communication infrastructure; it realizes the data collection of physical targets employing technologies such as WSNs, radio-frequency identification (RFID), ZigBee or near-field

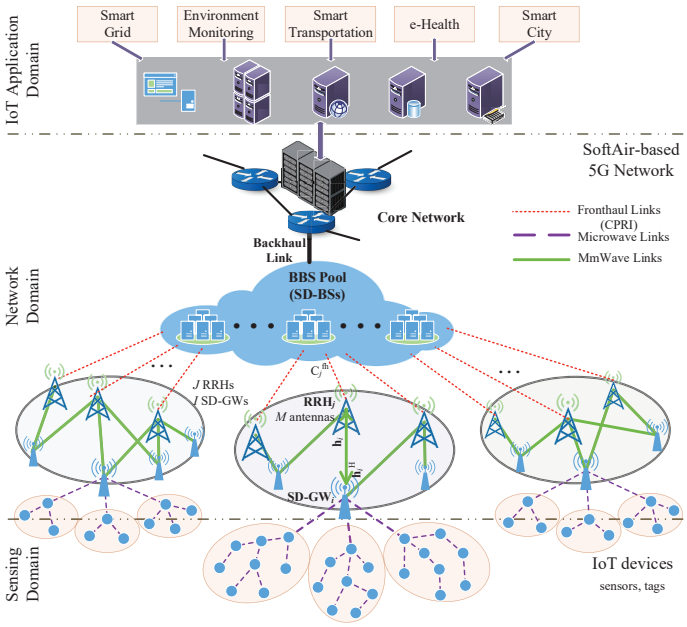


Figure 4.1: SoftAir [3] network architecture for 5G IoT communication.

communication (NFC). The network domain builds on SoftAir; it aims at transferring the data collected from the sensing domain to the remote destination in the application domain. Finally, the application domain is responsible for data processing and the provision of a wide variety of applications and services.

It is worth noting that a relevant architectural component is the SD-GW, that lies between the sensing and network domain. Besides alleviating the high traffic bursts imposed by sporadic communications from a myriad of heterogeneous IoT devices, the SD-GW aggregates the data from the IoT devices clustered geographically forming several WSNs and provides them access to the Internet using 5G wireless. In that sense, the SD-GW comprises two interfaces: northbound and southbound. The former communicates with the SoftAir system, whereas the latter interconnects the devices inside the

cluster, i.e., the SD-GW implements protocols that support co-existence of diverse wireless interfaces, such as intelligent management of interference and distributed management of channel allocation/medium access.

4.3.1 System model

Considering the SoftAir domain of the above architecture, the system consists of a set $\mathcal{I} = \{1, \ldots, I\}$ of SD-GWs that provide connectivity to several WSNs clustered geographically. Inside the WSN, the set $\mathcal{K} = \{1, \ldots, K\}$ of nodes communicate with neighboring devices by 6LoWPAN; we assume that the ith SD-GW acts as cluster head of the set $\mathcal{K}$ of nodes, and relays the data it receives to the SD-RAN. Then, the set $\mathcal{I}$ of SD-GWs is served by a set $\mathcal{J} = \{1, \ldots, J\}$ of associated RRHs. All the RRHs are connected to the BBS pool $\mathcal{B}$ via fronthaul links, where the jth fronthaul link between the $j \in \mathcal{J}$ RRH and the $b \in \mathcal{B}$ BBS has a predetermined capacity C_j^{fh}. Note that by using low-latency high-bandwidth fronthaul links, the software-defined architecture implements an accurate, high-resolution synchronization among RRHs and enables flexible and tangible RRH coordinations. The associations between RRHs and SD-GWs can be determined based on the distance or channel gain from RRHs to each SD-GW. The RRHs are equipped with an array of M antennas and communicate with the single antenna SD-GWs through mmWave links. One RRH can serve a number of SD-GWs; the jth RRH that is assigned to serve the ith SD-GW will receive the SD-GW's processed base band signal from the BBS pool. Then, the RRH converts and transmits the corresponding RF signal using a suitable designed pre-coding vector as detailed in Section 4.4.2.

In the SoftAir domain shown in Fig. 4.1, the SD-GW incorporates a local controller, and has the key role of being a concentrator of several sensor nodes for both control and user planes. It possesses the necessary knowledge to orchestrate the sensors such as network topology, link qualities, and application requirements. Besides performing conversions to communicate between different standards, the SD-GW performs the optimization frame-

work, and can make decisions such as the choice of network parameters and protocols. The network application can then be modified by simply changing the forwarding rules at the local controller, which then propagates the changes to sensors.

4.4 Heterogeneous cross-layer solution for software-defined gateway

In the following, we develop a cross-layer optimization framework that integrates the sensing domain and the SD-RAN of the SoftAir system, allowing coordination, interaction, and joint optimization of protocols crossing different layers. We elaborate the communication functionalities for both the sensing and SD-RAN domain in Section 4.4.1 and Section 4.4.2, respectively. Then, a centralized optimization framework to jointly control the parameters is formulated in Section 4.4.3 to ultimately reach an optimum configuration according to an application-dependent objective function. Finally, the protocol operation at the SD-GW is detailed in Section 4.4.4.

4.4.1 IoT & WSN network

In this section, we describe the parameters and communication functionalities at the physical layer (channel, modulation), link layer (channel coding, MAC), and network layer (addressing, routing) for the nodes in the sensing domain.

Physical layer functionalities

At the physical layer, the nodes follow the frequency spectrum allocation according to the IEEE 802.15.4 standard [65]; they might have different maximum transmission power and can select different modulation schemes. We use the log-normal channel model, which has been experimentally shown to model the low power communication in WSN accurately [66]. In this

model, the total path-loss in dB is given by $l^{\text{WSN}}(d_i)[\text{dB}] = l^{\text{WSN}}(d_0) + 10\bar{n}\log_{10}(d_i/d_0) + \eta$ for $d_i \geq d_0$, where d_i is the transmitter-receiver distance; d_0 is a reference distance; $\bar{n}$ is the path-loss exponent for a particular frequency band or environment; $\eta \sim \mathcal{N}(0, \sigma)$ is the large-scale shadow factor in dB; and $l^{\text{WSN}}(d_0) = 10\log_{10}((4\pi f d_0)/\text{c})^2$ is the path-loss at a reference distance, $d_0 = 1\,\text{m}$, for a given center frequency, $f \in \{800, 900, 2400\}\,\text{MHz}$; $\text{c} = 3 \times 10^8\,\text{ms}^{-1}$ is the speed of light; and $\bar{n} = 2$. The signal-to-noise ratio (SNR) at a distance d_i in the receiver, $\omega(d_i)$, is given by $\omega(d_i)[\text{dB}] = P_k^{\text{tx}} - l^{\text{WSN}}(d_i) - P^{\text{noise}}$, where P_k^{tx} [dBm] is the output power of the transmitter, and P^{noise} [dBm] denotes the total noise power at the receiver.

The transmission power and modulation have a direct impact on the bit error rate (BER). Given the link i, the BER Ψ_i is determined as a function of the adopted modulation technique, $mod_i \in \mathcal{M}$, and the SNR, $\omega(d_i)$, as

$$\Psi_i = \Psi(\omega(d_i), mod_i). \tag{4.1}$$

Note that $\Psi(\cdot)$ is well-known for standard modulations. In the sensing domain, we consider simple modulation schemes following the IEEE 802.15.4 standard [65], such as BPSK and OQPSK, which are suitable for energy-limited WSNs.

Link layer functionalities

Concerning the channel coding scheme, we advocate for the use of a hybrid automation repeat request (HARQ) error control scheme [66, 67] that results from the combination of forward error correction (FEC) codes for poor quality channel conditions (i.e., $\omega(d_i)$ low) as well as the merits of automation repeat request (ARQ) when the channel conditions are good (i.e., $\omega(d_i)$ high). Initially, an uncoded or lightly coded packet is transmitted; if the received packet has more errors than those that can be corrected by the chosen FEC code, a more robust FEC code is chosen. We consider block codes due to their energy efficiency and lower complexity compared to convolutional codes (CCs). For the link i, $cod_i \in \mathcal{C}$ denotes the adopted coding scheme with

coding rate RC_i. As far as the BCH$(bl; pl; ce)$ code with rate $RC_i = pl/bl$ is concerned, bl, pl, and ce denote block length, payload length, and the error correcting capability of FEC code in bits, respectively, and $ce < bl$. Given the BER as a function of the SNR and modulation adopted, see $\Psi_i(\cdot)$ in (4.1), the block error rate, Ψ_i^{block}, becomes $\Psi_i^{\text{block}} = \sum_{j=ce+1}^{bl} \binom{bl}{j} \Psi_i(\cdot)^j (1 - \Psi_i(\cdot))^{bl-j}$. Additionally, with $\wp$ bits being the packet length, the packet error rate (PER) Φ_i is calculated as follows

$$\Phi_i = 1 - (1 - \Psi_i^{\text{block}})^{\lceil \frac{\wp}{pl} \rceil}, \tag{4.2}$$

which is approximated as $\lceil \frac{\wp}{pl} \rceil \Psi_i^{\text{block}}$ when Ψ_i^{block} is small. As a result, in each transmission, the initial packet is either coded with a BCH$(128; 106; 3)$ code or not coded to reduce the Φ_i without drastically sacrificing the transmission data rate. If the first transmission fails, i.e., the number of errors is larger than the maximum number of bits that can be corrected, a more robust FEC code is used for the re-transmitted packet [e.g., BCH$(128; 78; 7)$] until the packet is successfully decoded or the maximum number of transmissions (including re-transmissions), $N_i^{\max}$, is reached. Using this HARQ error control scheme, the overall PER over link i is given by

$$\Phi_i^{\text{Rtx}} = \Upsilon(\Phi_i^{\text{uncoded}}, N_i^{\text{Tx-ub}}, ce), \tag{4.3}$$

where $\Upsilon(\cdot)$ is a function that relates Φ_i after HARQ error control scheme, Φ_i^{Rtx}, with the uncoded PER over link i, Φ_i^{uncoded}, which is derived next [see (4.4)] considering the data storage capacity of nodes; $N_i^{\text{Tx-ub}}$ is the upper-bound for the number of transmissions of a packet with correctly decoding over link i computed as $N_i^{\text{Tx-ub}} = (1 - \Phi_i^{\text{uncoded}})^{-1}$. Additionally, we take into account the data storage capacity of the sensor nodes, mem, that is related to the probability of discarding a packet at link i, $\mathbb{P}_i^{\text{pkt-dropout}}$, due to the fact that it can not be queued at the transmitter or at receiver. We define this probability as $\mathbb{P}_i^{\text{pkt-dropout}} = \Gamma(mem_k, F_k)$, where $\Gamma(\cdot)$ is a function that relates the maximum number of packets, mem_k, that can be queued at the transmitter or receiver and the total local traffic (own and relayed), F_k. For instance, assuming Poisson traffic, the transmitter and receiver can be modeled as a

single server queue with mem_k buffer size and $(DR_i \cdot RC_i)/\wp$ [pkt/s] service rate, where DR_i [kbps] is the data rate transmission of link i. With these parameters, we determine the uncoded PER over link i, Φ_i^{uncoded}, as follows

$$\Phi_i^{\text{uncoded}} = (1 - \mathbb{P}_i^{\text{pkt-dropout}})[1 - (1 - \Psi_i(\cdot))^{\wp})]. \tag{4.4}$$

Regarding the MAC functionality, we consider a variation of sleep MAC (SMAC) and carrier sense multiple access with collision avoidance (CSMA/CA) for addressing energy efficiency and scalability. On the one hand, as sensor nodes are likely to be battery powered, we adopt the idea of SMAC in which sensor nodes periodically listen and sleep [68] so that the network lifetime of these nodes is prolonged. On the other hand, with CSMA/CA, a node attempts to reserve the channel by using request to send/clear to send handshake after it sees the channel idle for an inter-frame space amount of time. If the node fails to reserve the medium, it switches to sleep mode to save energy and waits for the next listening cycle. This medium access method can eliminate the interference drastically if the carrier sensing is properly performed. Note that, if a reservation-based protocol is used, data packet collisions will not occur. Hence, the hybrid of SMAC and CSMA/CA MAC protocol can save energy as well as reduce the interference among the sensor nodes avoiding the degradation of both BER and PER [67]. In our framework, the duration of listen and sleep cycles (T^{listen}, $T^{\text{sleep}} = 9 \times T^{\text{listen}}$ for a 10% duty cycle) is adaptive to the QoS requirements and they are set the same for all nodes in one cluster. Note that, the longer the sleep duration is, the lower the idle energy consumption, but the longer the end-to-end delay. We consider this duration parameter in the MAC protocol to interplay with physical layer parameters in the proposed cross-layer framework.

Network layer functionality

The IoT is expected to have an incredibly high number of *things*, and each of them should be retrievable with a unique IP address. Thus, we advocate for the use of IPv6 addressing in our framework, consistently with 6LoWPAN. A

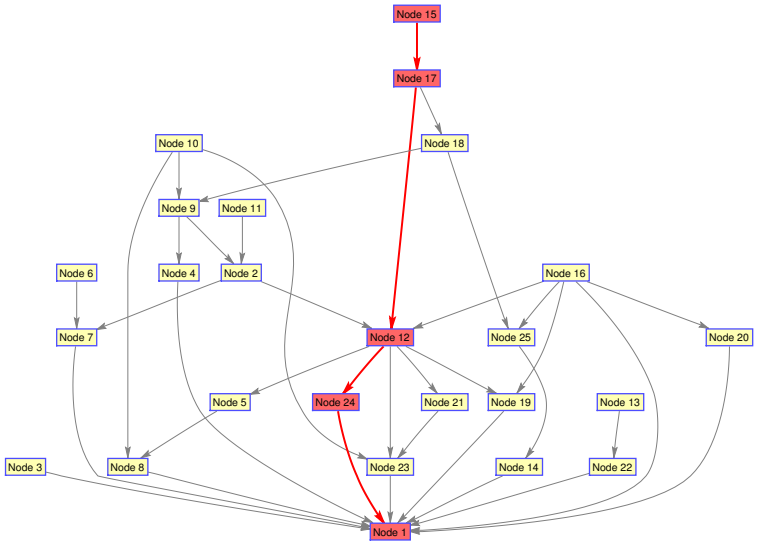


Figure 4.2: Directed Acyclic Graph (DAG) for a WSN consisting of 25 nodes randomly deployed and the optimal path (red color) from source to destination.

packet with fixed size ($\wp$ bits) is selected and used for all the links throughout a given path; the packet size is computed as $\wp = pl + h + ce$, where pl is the payload (data) length, h is the header length, and ce is the FEC redundancy length. We use the routing protocol for low power and lossy networks (RPL) for selecting the multi-hop paths that data packets follow from the source to reach the destination. RPL is a distance vector routing protocol that leaves the process of route selection to an external mechanism called *objective function*. RPL is based on the topological concept of destination oriented DAG (DODAG). The DODAG refers to a directed acyclic graph (DAG) with a single root as shown in Fig. 4.2; the costs associated with each directional link are derived accordingly to be consistent with the objective function (detailed in Section 4.4.3), the constraints, and the hardware capabilities of the *things* so that the optimal path from the source to destination is provided.

4.4.2 5G radio access network: SoftAir

Following the network model detailed in Section 4.3.1, we formulate the sum-rate optimization in the SoftAir SD-RAN, which jointly optimizes associations between RRHs and SD-GW that use mmWave transmissions, and RRHs' beamforming weights to maximize the SD-GW sum-rate while guaranteeing QoS and system-level constraints. We consider a short frame structure [69, 70] where time is discretized into frames, each frame has duration of T^{frame} symbols. We allocate τ_{ul} symbols for uplink transmission, and τ_{dl} symbols for downlink transmission.

Association scheme

Let $\mathcal{J} = \{1, \ldots, J\}$ and $\mathcal{I} = \{1, \ldots, I\}$ denote the set of RRHs and SD-GWs in the SoftAir system, respectively. Suppose that each SD-GW is served by a specific group of associated RRHs, and a RRH can serve multiple SD-GWs at the same time. To express the association status between RRHs and SD-GWs, we introduce the following binary variables as the indicators. Concretely, RRHs can be active to serve SD-GWs or shutdown to save the energy consumption, let $\{q_j, j \in \mathcal{J}\}$ denote the activity of RRHs as

$$q_j = \mathbb{I}[\text{the } j\text{th RRH is in active mode}]; \tag{4.5}$$

let $\{g_{ij}, i \in \mathcal{I}, j \in \mathcal{J}\}$ denote the association between RRHs and SD-GW as

$$g_{ij} = \mathbb{I}[\text{the } i\text{th SD-GW is served by the } j\text{th RRH}]; \tag{4.6}$$

furthermore, to characterize the group (cluster) of serving RRHs, let $\{N_{ij}, i \in \mathcal{I}, j \in \mathcal{J}\}$ be the clustering indicator as

$$N_{ij} = \mathbb{I}[(i, j) \in \mathcal{L}], \tag{4.7}$$

where $\mathbb{I}[x]$ is the indicator function, $\mathcal{L} = \{(i, j) \mid i \in \mathcal{I}, j \in \mathcal{N}_i\}$ denotes the predetermined set of feasible association, and $\mathcal{N}_i$ denotes the set of near RRHs for the ith SD-GW which can be determined based on the distance or channel gain from RRHs to each SD-GW.

Millimeter-wave communication

We introduce the link budget for mmWave communication between the ith SD-GW and jth RRH. Particularly we detail the path-loss, l_i, channel vector, $\mathbf{h}_i$, and beamforming gain, G_i^{BF}, for deriving both the achievable uplink and downlink data rates.

Path-Loss: Considering the peculiarities of mmWave propagation, the path-loss for a mmWave communication link i, l_i, can be modeled with three link-states: outage (l_{iO}), LoS (l_{iL}) or NLoS (l_{iN}) [71]. We formulate the path-loss with respect to these three states as follows

$$\begin{aligned} l_{iO} &= 0, \\ l_{iL} &= (\alpha_L d_i)^{-\beta_L}, \\ l_{iN} &= (\alpha_N d_i)^{-\beta_N}, \end{aligned} \tag{4.8}$$

where α_L (α_N) can be interpreted as the path-loss of the LoS (NoS) link at 1 [m] distance, and β_L (β_N) denotes the path-loss exponent of the LoS (NLoS) link. From experimental results [71], β_N value (can be up to 4) is normally higher than β_L value (i.e., 2). Then, each link-state is formulated by the channel state probabilities $\mathbb{P}_O$, $\mathbb{P}_L$, and $\mathbb{P}_N$, respectively, as

$$\begin{aligned} \mathbb{P}_O &= \max(0, 1 - \gamma_O e^{-\delta_O d_i}); \\ \mathbb{P}_L &= (1 - \mathbb{P}_O)\gamma_L e^{-\delta_L d_i}; \\ \mathbb{P}_N &= (1 - \mathbb{P}_O)(1 - \gamma_L e^{-\delta_L d_i}), \end{aligned} \tag{4.9}$$

where d_i denotes the transmitter-receiver distance; the parameters γ_L (γ_O) and δ_L (δ_O) depend on both the propagation scenario and the considered carrier frequency [72]. Thus, the corresponding path-loss component of the channel is modeled as

$$\begin{aligned} l_i =& \mathbb{I}[U < \mathbb{P}_L(d_i)] l_{iL} + \mathbb{I}[\mathbb{P}_L(d_i) \leq U < (\mathbb{P}_L(d_i) + \mathbb{P}_N(d_i))] l_{iN} \\ &+ \mathbb{I}[(\mathbb{P}_L(d_i) + \mathbb{P}_N(d_i)) \leq U \leq 1] l_{iO}, \end{aligned} \tag{4.10}$$

where $U \sim \mathcal{U}[0, 1]$ is a uniform random variable. For computing the path-loss model, we use the parameter values at 73 GHz as in [73, Table I].

Channel vector: Given that the blockage information is not entirely feasible, we exploit the stochastic geometry analysis for modeling the mmWave channel vector [73]. Specifically, we model the channel vector as $\mathbf{h}_i = \sqrt{l_i \boldsymbol{\beta}_i}\, \boldsymbol{\xi}_i \in \mathbb{C}^{M,1}$, where l_i is the large-scale path-loss in power of the mmWave communication link i (which might also include log-normal shadowing), $\boldsymbol{\beta}_i \in \mathbb{C}^{M,M}$ is the co-variance matrix for antenna correlations in small-scale fading, and $\boldsymbol{\xi}_i \in \mathbb{C}^{M,1}$ is a Gaussian vector with the zero-mean circularly symmetric Gaussian noise distribution $\mathcal{CN}(0, \mathbf{I}_M)$ for the fast-fading.

Beamforming: To ensure an acceptable range of the communication in the multi-antenna mmWave transmissions, we introduce the precode vectors, i.e., beamforming weights at the RRHs, where the weight vector $\mathbf{w}_i \in \mathbb{C}^{M,1}$ is the linear downlink beamforming vector at the jth RRH corresponding to the ith SD-GW. The beamforming gain is given as $G_i^{\mathrm{BF}} = \mathbf{w}_i^{\mathsf{H}} \boldsymbol{\beta}_i \mathbf{w}_i$, with $\boldsymbol{\beta}_i$ being the covariance matrix of the channel response vector $\mathbf{h}_i$. In the case where the fading is fully correlated between the antennas, the matched filtering pre-coding method is exploited as $\boldsymbol{\beta}_i = \mathbf{h}_i^{\mathsf{H}} \mathbf{h}_i$ and $\mathbf{w}_i = \mathbf{h}_i / \|\mathbf{h}_i\|$; therefore, $G_i^{\mathrm{BF}} = \|\mathbf{h}_i\|^2$.

Achievable uplink rate

Following the above multi-antenna mmWave transmission characterization over a link i, the received base-band signal vector $\mathbf{y} \in \mathbb{C}^{M,1}$ at the BBS at a given instant reads $\mathbf{y}^{\mathrm{ul}} = \sqrt{P^{\mathrm{ul}}}\, \mathbf{H}\mathbf{x}^{\mathrm{ul}} + \eta^{\mathrm{ul}}$, where each element of the received signal vector corresponds to a BBS antenna, $\mathbf{H} = [\mathbf{h}_1 \cdots \mathbf{h}_I] \in \mathbb{C}^{M,I}$, $\mathbf{h}_i \in \mathbb{C}^{M,1}$ denotes the mmWave channel corresponding to the ith SD-GW, $\mathbf{x} = [x_1 \cdots x_I]^{\mathsf{T}}$ denotes the I×1 vector containing the transmitted signals from all the SD-GWs, P^{ul} is the average transmit power of each SD-GW, and $\eta^{\mathrm{ul}} \sim \mathcal{CN}(0, \sigma)$ is the zero-mean circularly symmetric Gaussian noise with the noise power σ^2.

Let $\mathbf{A}$ be the M×I linear detection matrix (which depends on the channel matrix $\mathbf{H}$) used by the BBS $b \in \mathcal{B}$ to separate the received signal into user streams. The BBS processes its received signal vector and obtains the

estimated channel matrix (assuming no estimation errors) by multiplying the detection matrix with the Hermitian-transpose of the linear receiver as $\tilde{\mathbf{y}}^{\mathrm{ul}} = \mathbf{A}^{\mathsf{H}}\mathbf{y}^{\mathrm{ul}} = \mathbf{A}^{\mathsf{H}}\mathbf{H}\mathbf{x} + \mathbf{A}^{\mathsf{H}}\eta^{\mathrm{ul}}$. The ith element of $\tilde{\mathbf{y}}^{\mathrm{ul}}$ can be written as $\tilde{y}_i^{\mathrm{ul}} = \sqrt{P_i^{\mathrm{ul}}}\mathbf{a}_i^{\mathsf{H}}\mathbf{H}\mathbf{x} + \mathbf{a}_i^{\mathsf{H}}\eta^{\mathrm{ul}}$, where $\mathbf{a}_i$ is the ith column of $\mathbf{A}$. By the elements multiplication, we further get $\tilde{y}_j^{\mathrm{ul}} = \sqrt{P_i^{\mathrm{ul}}}\mathbf{a}_i^{\mathsf{H}}\mathbf{h}_i x_i + \sum_{k=1,k\neq i}^{I}\sqrt{P_k^{\mathrm{ul}}}\mathbf{a}_i^{\mathsf{H}}\mathbf{h}_i x_k + \mathbf{a}_i^{\mathsf{H}}\eta^{\mathrm{ul}}$, where x_i denotes the ith element of $\mathbf{x}$ and $\mathbf{h}_i$ is the ith column of $\mathbf{H}$. Then, the signal-to-interference-plus-noise ratio (SINR) achieved by the ith SD-GW, γ_i^{ul}, is

$$\gamma_i^{\mathrm{ul}} = P_i^{\mathrm{ul}}|\mathbf{a}_i^{\mathsf{H}}\mathbf{h}_i|^2 / (\textstyle\sum_{k=1,k\neq i}^{I} P_k^{\mathrm{ul}}|\mathbf{a}_i^{\mathsf{H}}\mathbf{h}_k|^2 + \|\mathbf{a}_i\|^2\sigma^2). \tag{4.11}$$

Assuming an ergodic channel [74], the achievable uplink rate of the ith SD-GW is given by $R_i^{\mathrm{ul}} = B\log_2(1+\gamma_i^{\mathrm{ul}})$, where B denotes the wireless transmission bandwidth. We define the uplink sum rate [bits/s/Hz] per cell considering the associations between RRHs and SD-GWs as follows

$$C^{\mathrm{ul}} = \textstyle\sum_{j=1}^{J}\sum_{i=1}^{I} g_{ij}\, N_{ij}\, R_i^{\mathrm{ul}}. \tag{4.12}$$

Achievable downlink rate

The received base band signal $y^{\mathrm{dl}} \in \mathbb{C}$ at the ith SD-GW is given as $y^{\mathrm{dl}} = \sqrt{P_j^{\mathrm{dl}}}\mathbf{h}_i^{\mathsf{H}}\mathbf{s} + \eta^{\mathrm{dl}}$, where $\mathbf{s} \in \mathbb{C}^{M,1}$ is the signal vector intended for the ith SD-GW with P_j^{dl} average power; $\eta^{\mathrm{dl}} \sim \mathcal{CN}(0,\sigma^2)$ is the receiver noise. We assume channel reciprocity, i.e., the downlink channel $\mathbf{h}_i^{\mathsf{H}}$ is the Hermitian transpose of the uplink channel $\mathbf{h}_i$. The transmit vector $\mathbf{s}$ is given as $\mathbf{s} = \sqrt{v}\sum_{i=1}^{I}\mathbf{w}_i x_i^{\mathrm{dl}} = \sqrt{v}\,\mathbf{W}\,\mathbf{x}^{\mathrm{dl}}$, where $\mathbf{W} = [\mathbf{w}_1 \cdots \mathbf{w}_I] \in \mathbb{C}^{M,I}$ is a pre-coding matrix (i.e. the network beamforming design) and $\mathbf{x}^{\mathrm{dl}} = [x_1 \cdots x_I]^{\mathsf{T}} \in \mathbb{C}^{I,1}$ contains the data symbols for the ith SD-GW. The parameter v normalizes the average transmit power per RRH to $\mathbb{E}[\frac{P_j^{\mathrm{dl}}}{I}\mathbf{s}^{\mathsf{H}}\mathbf{s}] = P_j^{\mathrm{dl}}$, i.e., $v = \left(\mathbb{E}\left[\frac{1}{I}\mathrm{tr}(\mathbf{W}\mathbf{W}^{\mathsf{H}})\right]\right)^{-1}$.

The associated SINR achieved by the ith SD-GW, γ_i^{dl}, is

$$\gamma_i^{\mathrm{dl}} = v|\mathbf{h}_i^{\mathsf{H}}\mathbf{w}_i|^2 / \left(\textstyle\sum_{k=1,k\neq i}^{I} v\,|\mathbf{h}_i^{\mathsf{H}}\mathbf{w}_k|^2 + \sigma^2\right). \tag{4.13}$$

Since the SD-GWs do not have any channel estimate, we provide an ergodic achievable rate based on the techniques developed in [74, Theorem 1] as $R_i^{\mathrm{dl}} = B_i(1-\kappa)\log_2(1+\gamma_i^{\mathrm{dl}})$, where B_i is the bandwidth allocated to the ith SD-GW, κ accounts for the spectral efficiency loss due to signaling at RRH. The downlink sum rate [bits/s/Hz] per cell considering the associations between RRHs and SD-GWs is

$$C^{\mathrm{dl}} = \sum_{j=1}^{J}\sum_{i=1}^{I} g_{ij}\, N_{ij}\, R_i^{\mathrm{dl}}. \tag{4.14}$$

4.4.3 Optimization framework

In the following, we elaborate the optimization framework for the sensing and SD-RAN domain.

IoT sensing domain

The IoT should provide services for applications with different QoS requirements, ranging from error-limited applications or minimum energy consumption applications to highly-delay-sensitive applications or any combination of them. Hence, we consider a multi-objective optimization problem which can simultaneously optimize multiple conflicting end-to-end objectives such as PER (Φ^{e2e}), delay (T^{e2e}), and energy consumption (E^{e2e}), subject to certain constraints.

We construct a single aggregate objective function which is defined by the weighted linear combination of each objective; we use w_{PER}, w_{E}, and w_{T} as the three weights for the end-to-end PER, energy consumption, and time delay objectives, respectively. As these three objectives differ in the units in which they are measured as well as their order of magnitude, we normalize each term and optimize their deviations with respect to some predefined utopia values (unattainable minimum values which are used to provide the non-dimensional objective functions and can be computed offline [67]).

Therefore, the overall objective function for WSN communication becomes

$$\text{minimize} \quad w_{\text{PER}}\left|\frac{\Phi^{\text{e2e}}}{\Phi^{\text{opt}}}-1\right|+w_{\text{E}}\left|\frac{E^{\text{e2e}}}{E^{\text{opt}}}-1\right|+w_{\text{T}}\left|\frac{T^{\text{e2e}}}{T^{\text{opt}}}-1\right|, \tag{4.15}$$

where $w_{\text{PER}}+w_{\text{E}}+w_{\text{T}}=1$; Φ^{opt}, E^{opt}, T^{opt} are the end-to-end PER, energy consumption, and delay utopia values for normalizing purposes, respectively. Note that (4.15) may target at different degrees of QoS requirements for various IoT applications by adapting the specific weight value (w_{PER}, w_{E}, or w_{T}) according to the application.

Statistical QoS guarantee: The higher transmission reliability associated with lower PER is crucial for almost all types of WSN. Also, having a bounded delay is especially important for real-time monitoring and applications with timing constraints. Aiming to support the distributed functionalities among sensors, in the following we form the per-node based constraints (i.e., for transmissions upon link i) of link reliability, delay, and energy.

Given the tolerable maximum end-to-end PER, Φ^{TH}, the corresponding reliability constraint is

$$\Phi^{\text{e2e}}=\left(1-(1-\Phi_i^{\text{Rtx}})^{N^{\text{hops}}}\right)\leq\Phi^{\text{TH}}, \tag{4.16}$$

where Φ_i^{Rtx} [see (4.3)] is the PER over link i with HARQ error control, and N^{hops} is the number of traversed hops for an incoming packet to node k.

Regarding the energy consumption, let E_k denote the energy consumed on the kth node, it is defined by the product of the packet size and the energy required for one bit as $E_k=\wp(2E_{\text{elec}}^{\text{bit}}+P_k^{\text{tx}}/F_k)$, where $E_{\text{elec}}^{\text{bit}}=E_{\text{elec}}^{\text{bit-Tx}}=E_{\text{elec}}^{\text{bit-Rx}}$ in Joule/bit is the distance-independent energy to transmit one bit; $E_{\text{elec}}^{\text{bit-Tx}}$ is the energy per bit needed by the transmitter electronics, and $E_{\text{elec}}^{\text{bit-Rx}}$ is the energy per bit utilized by the receiver electronics; P_k^{tx} and F_k are the transmission power and the total local traffic at the kth node, respectively. Restricted by the constraint E^{TH}, the overall energy consumption over the entire path is computed by

$$E^{\text{e2e}}=\textstyle\sum_{k=1}^{N^{\text{hops}}}E_k\leq E^{\text{TH}}. \tag{4.17}$$

Finally, restricted by the maximum end-to-end delay T^{TH}, the statistical delay guarantee is modeled as the probability that a packet is delivered under the deadline should be at least φ as follows

$$\mathbb{P}(T^{\text{e2e}} \leq T^{\text{TH}}) \geq \varphi. \tag{4.18}$$

The end-to-end delay, T^{e2e}, is calculated as $T^{\text{e2e}} = \sum_{i=1}^{N^{\text{hops}}} (T_i^{\text{queuing}} + T_i)$, where T_i^{queuing} is the queuing delay at link i and T_i is the delay at link i excluding the queuing delay. T_i is composed of the time for handshake $T_i^{\text{handshake}}$, time for data transmission T_i^{data}, timeout delay T_i^{timeout}, time for acknowledgment T_i^{ack}, sleep time T^{sleep}, and the signal processing time T_i^{DSP}, and is calculated as

$$\begin{aligned} T_i \leq & (T_i^{\text{handshake}} + T_i^{\text{data}} + T_i^{\text{timeout}})(N_i^{\text{Tx-ub}} - 1) \\ & + (T_i^{\text{handshake}} + T_i^{\text{data}} + T_i^{\text{ack}}) + T^{\text{sleep}} + T_i^{\text{DSP}}. \end{aligned} \tag{4.19}$$

Note that the queuing delay is determined by many factors such as current traffic, other nodes' behavior or hardware status. Therefore, the overall end-to-end delay is modeled, by applying the central limit theorem, as a Gaussian random variable

$$T^{\text{e2e}} \sim \mathcal{N}\left(\sum_{i=1}^{N^{\text{hops}}} \left(T_i + \mathbb{E}[T_i^{\text{queuing}}] \right), \left(\sum_{i=1}^{N^{\text{hops}}} \mathbb{V}[(T_i^{\text{queuing}})] \right)^{1/2} \right). \tag{4.20}$$

Then, the end-to-end delay constraint (4.18) is transformed into

$$\sum_{i=1}^{N^{\text{hops}}} \left(T_i + \mathbb{E}[T_i^{\text{queuing}}] \right) + \phi^{-1}(\varphi)(\sum_{i=1}^{N^{\text{hops}}} \mathbb{V}[(T_i^{\text{queuing}})])^{1/2} \leq T^{\text{TH}}, \tag{4.21}$$

where $\phi(\cdot)$ is the cumulative distribution function of $\mathcal{N}(0,1)$. Finally, the end-to-end throughput, G^{e2e}, is inversely proportional to the end-to-end delay as $G^{\text{e2e}} = pl/T^{\text{e2e}}$.

SoftAir SD-RAN domain

As IoT applications demand services with different rate requirements, we formulate these requirements in terms of SINR coverage and achieved sum-rate per cell at the SD-RAN. Given ϑ as the minimum tolerable SINR over a

link i, the SINR constraints of SD-GWs can be formulated as

$$\gamma_i \geq \vartheta, \forall i \in \mathcal{I}, \tag{4.22}$$

where γ_i is computed by either (4.11) or (4.13) in case of uplink or downlink transmission, respectively. From the association scheme, we can obtain the equality $q_j = 1 - \prod_{j=1}^{J} (1 - g_{ij} N_{ij}), \forall j \in J$ and the following sets of association constraints between RRHs and SD-GW:

$$q_j \geq g_{ij}\, N_{ij}, \forall i \in \mathcal{I}, j \in \mathcal{J}; \tag{4.23}$$

$$\sum_{j=1}^{J} g_{ij}\, N_{ij} \geq 1, \forall i \in \mathcal{I}, \tag{4.24}$$

where (4.23) implies that a RRH is in active mode if it is associated with at least one SD-GW whereas (4.24) ensures that each SD-GW is served by at least one RRH. On the other hand, given the pre-coding vector at the jth RRH for the ith SD-GW, the transmitter power used by this RRH to serve the ith SD-GW is $\mathbf{w}_i^{\mathsf{H}} \mathbf{w}_i$ [75]. Let $P_j^{\text{r-max}}$ denote the maximum power of the jth RRH, we impose the constraints on RRHs' downlink beamforming weights as follows

$$\sum_{i=1}^{I} \mathbf{w}_i^{\mathsf{H}} \mathbf{w}_i \leq q_j\, P_j^{\text{r-max}}, \forall j \in \mathcal{J}; \tag{4.25}$$

$$\mathbf{w}_i^{\mathsf{H}} \mathbf{w}_i \leq g_{ij}\, N_{ij}\, P_j^{\text{r-max}}, \forall i \in \mathcal{I}, j \in \mathcal{J}, \tag{4.26}$$

where (4.25) limits the total transmit power of RRHs and (4.26) ensures that the transmit power from the jth RRH to the ith SD-GW is set to zero if there is no association between them. Furthermore, by only allowing the links in $\mathcal{L}$ (see Section 4.4.2) we set the beamforming weights of mmWave communication links as

$$\mathbf{w}_i^{\mathsf{H}} \mathbf{w}_i = 0 \text{ if } N_{ij} = 0, \forall i \in \mathcal{I}, j \in \mathcal{J}, \tag{4.27}$$

so that we reduce all possible links between J RRHs and I SD-GWs to $|\mathcal{L}|$ links (given that $|\mathcal{L}| \ll JI$), which in turns dramatically shrinks the possible solution sets of precoding vectors for lower computation complexity [73, 75].

Additionally, the per-fronthaul capacity constraints (neglecting the fronthaul capacity consumption for transferring compressed beamforming vector) are formulated as follows

$$C \leq C_j^{\text{fh}}, \forall j \in \mathcal{J}, \tag{4.28}$$

where C is computed by (4.12) in uplink transmission or by (4.14) in downlink transmission. This indicates that the total data rate transmitted at the jth RRH should be less or equal to the rate forwarded by the jth fronthaul link.

We aim to maximize the total achievable uplink/downlink data rates at SD-GWs; the overall objective function for the SD-RAN becomes

$$\text{maximize} \quad C = \textstyle\sum_{i=1}^{I} R_i, \tag{4.29}$$

where R_i is computed based on the communication direction: uplink or downlink (see Section 4.4.2).

Statistical QoS guarantee: To ensure low transmission delay, the size of each packet, $\wp$ bits, is small enough such that it can be transmitted within one uplink phase; the transmission time interval (TTI) is the same as the frame duration, T^{frame}, and $T^{\text{frame}} \ll T^{\text{TH}}$. Thus, the uplink (downlink) transmission can be finished within the duration of τ^{ul} (τ^{dl}). Furthermore, the expected queuing delay for the packets at the SD-GW should be bounded as

$$\mathbb{E}[T_i^{\text{queuing}}] \leq T^{\text{TH}} - T^{\text{frame}}. \tag{4.30}$$

Then, to guarantee the stringent QoS requirements for all SD-GW $i \in \mathcal{I}$, our cross-layer design satisfies that: *(i)* the probability that the queuing delay is larger than $(T^{\text{TH}} - T^{\text{frame}})$ is smaller than a predefined violation probability D^{TH}, i.e., $\mathbb{P}(T_i^{\text{queuing}} > (T^{\text{TH}} - T^{\text{frame}})) < D^{\text{TH}}$; *(ii)* with finite BCH codes, the transmission of each packet is finished within one frame with a small error probability, i.e., $\Phi_i^{\text{e2e}} \leq \Phi^{\text{TH}}$; *(iii)* to guarantee the end-to-end delay and its reliability with finite transmit power, the packet dropout probability, $\mathbb{P}_i^{\text{pkt-dropout}}$ is smaller than a predefined violation probability Q^{TH}; *(iv)* the probability that the SINR coverage is smaller than ϑ is smaller than a predefined violation probability ϑ^{TH}, i.e., $\mathbb{P}(\gamma_i < \vartheta) < \vartheta^{\text{TH}}$. Finally, the end-to-end

Table 4.1: Heterogeneous Cross-Layer Optimization Framework

Inputs:

Sensor domain:	Φ^{opt}, E^{opt}, T^{opt}, Φ^{TH}, E^{TH}, T^{TH}, D^{TH}, Q^{TH}, h, $N_k^{\max}$, mem_k, $\forall k \in \mathcal{K}$
Cellular domain:	$P_j^{\text{r-max}}$, C_j^{fh}, ϑ^{TH}, Ω, T^{frame}, $\forall j \in \mathcal{J}$
Compute (offline):	w_{PER}, w_{E}, w_{T}, $N_k^{\text{Tx-ub}}$, F_k, $\forall k \in \mathcal{K}$

Find:

Sensor domain:	P_k^{tx}, mod_i, cod_i, $\wp$, T^{listen}, $\forall k \in \mathcal{K}, i \in \mathcal{I}$
Cellular domain:	P_i^{ul}, P_j^{dl}, q_j, g_{ij}, $\mathbf{w}_i$, $\forall i \in \mathcal{I}, j \in \mathcal{J}$

Objectives:

$$\text{minimize} \quad w_{\text{PER}}\left|\frac{\Phi^{\text{e2e}}}{\Phi^{\text{opt}}} - 1\right| + w_{\text{E}}\left|\frac{E^{\text{e2e}}}{E^{\text{opt}}} - 1\right| + w_{\text{T}}\left|\frac{T^{\text{e2e}}}{T^{\text{opt}}} - 1\right| \quad (4.15)$$

$$\text{maximize} \quad C = \sum_{i=1}^{I} R_i. \quad (4.29)$$

Subject to:

Packet error rate constraints:	(4.16), (4.4), (4.3).
Energy consumption constraint:	(4.17).
Delay constraints:	(4.30), (4.21).
SINR constraints:	(4.22); (4.11) UL, (4.13) DL.
Association constraints:	(4.24), (4.23).
Beamforming weights constraints:	(4.27), (4.26), (4.25).
Per-fronthaul capacity constraint:	(4.28).
System reliability constraint:	(4.31).

system reliability is controlled by

$$1 - (1 - D^{\text{TH}})(1 - \Phi^{\text{TH}})(1 - Q^{\text{TH}})(1 - \vartheta^{\text{TH}}) \leq \Omega, \quad (4.31)$$

where Ω dictates the overall reliability requirement. The entire formulation of the jointly cross-layer optimization for SD-GWs is summarized in Table 4.1.

4.4.4 Protocol operation

The SD-GWs possess the necessary knowledge (e.g., network topology, link qualities) to orchestrate sensors at the southbound interface and the application requirements at the northbound interface. Therefore, they are able to receive IoT data traffic from the sensing devices and forward this traffic to the SoftAir SD-RAN. Depending on the communication direction, each SD-GW will either perform protocol conversions in such a way it can forward the data to the SoftAir system with the maximum achievable rate or forward the data to the WSN meeting the application QoS requirements by performing the optimization framework.

The SD-GW first builds the hierarchical topology (DODAG) that specifies a route from each node to itself using a DODAG Information Object (DIO) message. Once a node receives a DIO, it can first, calculate its rank, then, choose a set of parent nodes (candidate nodes where data can be forwarded) and finally, send a new DIO message to inform other neighbors. The rank is an integer that grows linearly from the SD-GW and identifies the position of the node about the SD-GW and other nodes in the network. The parents must have a rank equal or lower than the node. Each node has a default path (i.e., preferred parent) but maintains a list of parents for resilience purposes, overhead reduction in case of link degradation, or for increasing performance. Hence, immediately after the reception of a DIO message, a node has an optimal path towards the SD-GW.

The optimal path is set according to the optimization framework detailed in Section 4.4.3. The SD-GW translates the application requirements into network QoS requirements and constructs the optimal architecture finding the optimal communication parameters on both the sensor and the SD-RAN domains and forwarding the data through the correspondent interface. The SD-GWs determine the routing paths at the local level, and at the inter-cluster level, cluster coordinators, who are elected by the transmission algorithm, facilitate the communication.

4.5 Performance evaluation

In this section, we present simulation results to assess and compare the performance of the proposed cross-layer design detailed in Section 4.4 with that of conventional layered protocol solutions, i.e., individual communication functionalities that do not share information and operate in separate layers. In all experiments, each point represents the average value of 10^5 samples. The overall reliability requirement is $\Omega = 10^{-5}$.

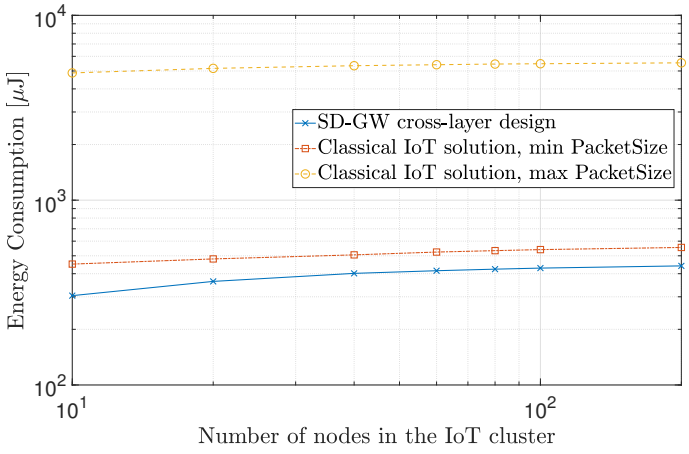
Following the system model (see Section 4.3.1), we design a set $\mathcal{J}$ of $J = 12$ associated RRHs in the SD-RAN, each RRH is equipped with $M = 4$ antennas; the coverage area of every RRH has a radius of 200 m. The channel vectors are generated according the mmWave communication characterization detailed in Section 4.4.2, where the three-state path-loss model with log-normal shadowing is considered; the carrier frequency is set at 73 GHz. The transmit power constraint for each RRH is $P_j^{\text{r-max}} = 45\,\text{dBm}$. Moreover, we assume that all the RRHs possess the same fronthaul capacity, i.e., $C_j^{\text{fh}} = 6\,\text{bps/Hz}$, since 64 QAM is the highest constellation supported by the network, and thus the maximum spectrum efficiency per data stream is 6 bps/Hz. The bandwidth of the wireless link is $B = 500\,\text{MHz}$. In the sensing domain, the set $\mathcal{J}$ of associated RRHs serve a set $\mathcal{I}$ of clustered sensor nodes. The ith cluster has one SD-GW as the cluster head. The maximum transmission power of each SD-GW is set at 23 dBm and thermal noise power is assumed to be $-101\,\text{dBm/Hz}$. The sensor nodes inside the cluster are randomly deployed; the concerned coverage area of each sensor node has a radius of 50 m. Each node uploads packets with rate 0.02 packets/TTI; the TTI has set to 0.1 ms. We compare the results when the QoS requirements are focused on end-to-end delay minimization and energy consumption minimization while the PER is constrained to be below $\Phi^{\text{TH}} = 10^{-6}$. The packet sizes that we consider in the simulations are $\wp = \{20, 40, 100, 133\}$ bytes.

We first examine the interactions between link layer and routing functionalities via end-to-end energy consumption and throughput performance of one IoT cluster at the sensing domain and compare the results of our de-

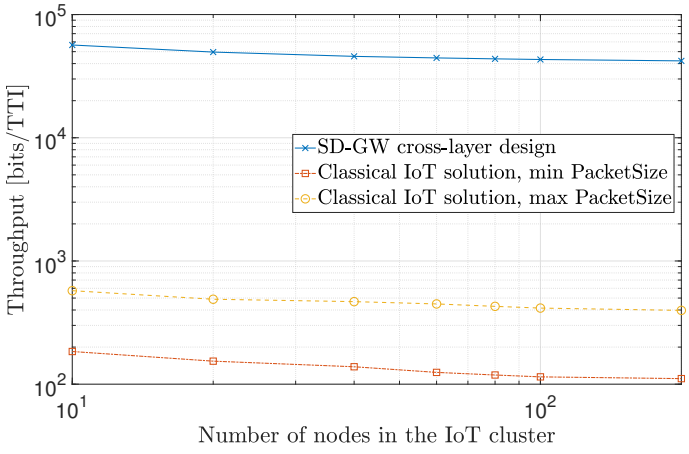
sign with that of a classical IoT communication solution. Then, we examine both the sum-rate and the achievable rate per SD-GW in the SD-RAN as the number of IoT clusters increases. For this, we consider that each cluster has a fixed number of sensor nodes and one SD-GW as cluster head. A layered protocol architecture is built for the comparison with our proposed design; its configuration is the following:

Classical IoT Solution (sensing domain: IEEE 802.15.4 + RPL; SD-RAN domain: conventional association schemes used in mmWave [72, 73]) In the sensing domain, this protocol configuration follows the frequency spectrum allocation according to the IEEE 802.15.4 standard at 2400 MHz (OQPSK modulation, 250 kbps transmission rate) and Sleep MAC + CSMA/CA for the PHY and link layer, respectively. In the NET layer, this protocol uses RPL and, for a fair comparison, the objective function is similar to that of our design (focused on the minimization of end-to-end delay and energy consumption while the PER is constraint to be below $\Phi^{\mathrm{TH}} = 10^{-6}$). In the SD-RAN, the following association schemes used in conventional mmWave communication are configured: (i) highest received power association and (ii) smallest path-loss association. Although the described layered protocol applies previously proposed functionalities more or less, it only considers its related layers without information sharing but with reasonable assumptions for the other layers.

Fig. 4.3a and Fig. 4.3b show, in logarithmic scale, the end-to-end energy consumption [μJ] and throughput [bits/TTI], respectively, as a function of the number of nodes in the cluster. In Fig. 4.3a we observe that the energy consumption of the SD-GW design is always lower than that of the classical IoT solution; on average, the energy savings of our solution ranges from 22.6% up to 92.5%. Also, we can observe that the energy consumption increases gradually as the number of nodes in the network increases since the higher node density essentially creates more paths for the data transmission. It is evident that using large packet sizes imply the highest levels in energy consumption. Fig. 4.3b shows the significant improvement of throughput with the SD-GW cross-layer design. The reason is that our solution selects the optimum path



(a)



(b)

Figure 4.3: IoT performance metrics vs. number of nodes in the network for the proposed design and the classical IoT solution. (a) Energy Consumption. (b) Throughput.

from the source to destination and the best parameter configuration for the device such as power, modulation, coding scheme, packet size.

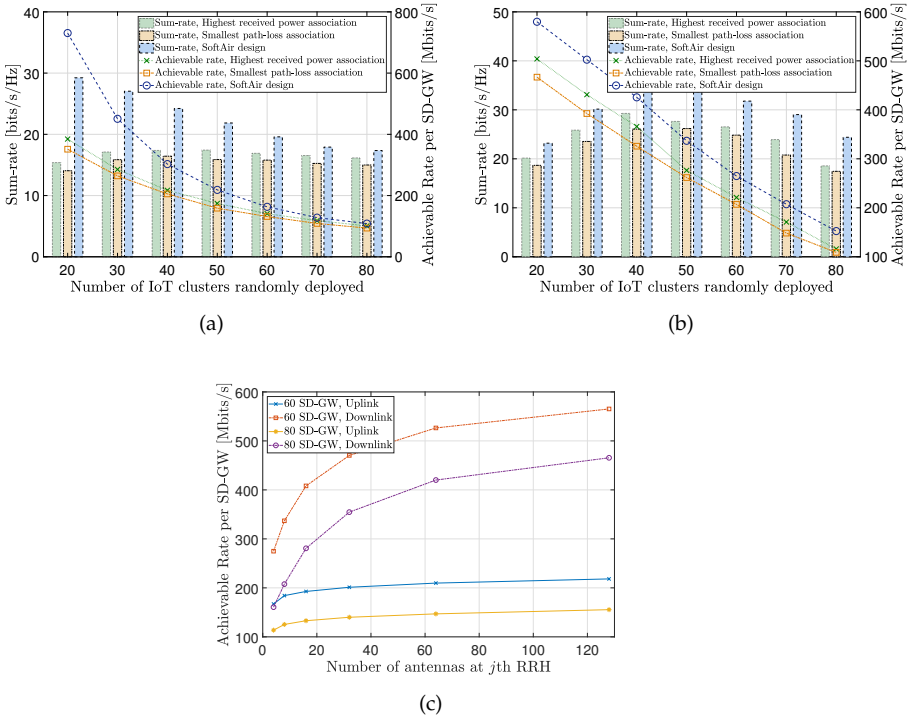


Figure 4.4: Sum-rate and achievable rate vs. number of SD-GWs deployed for the SoftAir design and conventional association schemes in mmWave; (a) upstream transmissions, (b) downstream transmissions. (c) Impact of increasing the number of antenna elements at RRHs on SD-GWs' achievable rate.

On the other hand, we analyze the achievable rate per SD-GW in the SD-RAN domain, the set $\mathcal{J}$ of associated RRHs serve different densities of IoT clusters; each IoT cluster has 100 nodes and one SD-GW as cluster head.

In upstream transmissions, Fig. 4.4a shows that our design outperforms conventional association schemes used in mmWave transmissions with coordinated multi-point. On average, the spectral efficiency of our design is 40%

higher than that of other association schemes and the achieved data-rate is up to 54% higher than that of conventional solutions. Regarding downstream transmissions, Fig. 4.4b shows that, on average, our design outperforms 26% conventional association schemes in terms of spectral efficiency. Furthermore, the spectral efficiency peaks at 34 bits/s/Hz; then, it slightly declines as the number of SD-GW increases.

Although the achievable data rate per SD-GW decreases with the increasing SD-GW density, our architecture can provide high data-rate for each SD-GW by increasing the number of antennas at the RRHs. This fact is shown in Fig. 4.4c for high densities of SD-GWs (60 and 80); specifically, with an antenna array of 128 elements at the RRHs and 80 SD-GWs deployed in such an area, our design can support each SD-GW with at least 450 [Mbits/s] rate in downlink and 150 [Mbits/s] rate in uplink through mmWave transmissions.

4.6 Highlights

We have presented a SoftAir architecture for providing IoT communication by exploiting a set of emerging features such as mmWave and SDN. Our solution brings significant system synergies by jointly optimizing functionalities in different communication layers for both IoTs and SD-RANs; SD-GWs are proposed to (i) explore the interactions of two-type networks, (ii) enable cross-layer solutions, and (iii) render efficient energy consumption and throughput in IoT, while maximizing the sum-rate at the SD-RAN for reliable IoT communication. Simulation results validate the superiority of our solutions that provide performance improvements in terms of energy savings, throughput, and spectral efficiency in comparison with conventional IoT solutions. It allows enormous and reliable IoT connectivity with high data rates at 5G SD-RANs.

Chapter 5

Performance Analysis of Wireless Cellular Networks based on Time-Scale Separation

5.1 Introduction

Nowadays, wireless communication networks incorporate sophisticated technology and algorithms to provide a wide range of services. In order to evaluate their performance and to understand the interactions among different components of these rather complex networks, the deployment of analytical models has become a common approach with multiple advantages. Accurate modeling of the wireless network events allows to determine performance parameters like the blocking probability, throughput, average transfer delay, and others [76,77].

The increasing complexity of wireless networks in terms of size, different configurations, and the interactions among types of traffic flows makes modeling more challenging. From the modeling perspective, we normally encounter two main common characteristics in continuous-time Markov chain (CTMC) models of wireless networks. First, the cardinality of the state-space of their CTMC is large. Second, the multiple types of traffic flows evolve at different time-scales.

While the first characteristic usually makes the exact solution of the CTMC computationally intractable, the second one allows us to apply specific solution approaches that exploit the time-scale separation to reduce the computational cost. We can structure the model into subsets of states by using the fact that transitions occur at a fast time-scale in the states belonging to the same subset, while transitions between subsets occur at a slower time-scale. Then, we can approximate the solution of the stationary probability distribution of the complete system by computing separately the stationary distribution of each subset, and then combining them to obtain the stationary distribution of the complete system. Once this is achieved, the performance metrics of the wireless network can be easily computed [78, 79].

The main contribution presented in this chapter is a new approximation method applicable to a wide range of time-scale separations, and whose accuracy can be improved by increasing the computational cost. The proposed method is based on an original iterative approach named absorbing Markov chain approximation (AMCA). We used it to analyze two different multiservice networks. One is a cognitive radio network (CRN) with two channel sets: one shared by primary and secondary users, and the other dedicated to the secondary users [80,81]. The other is an integrated service network (ISN), where a single base station serves real-time and non-real-time traffic [82, 83]. We will refer to these two networks as the *test networks*.

We carry out two types of analysis in the test networks. First, we evaluate the behavior of AMCA at different time-scale separations. Second, we study the trade-off between accuracy and computational cost. We compare the performance of AMCA with that of quasi-stationary approximation (QSA), generalized quasi-stationary approximation (GQSA), and a classical iterative method named iterative aggregation/disaggregation (IAD), which is particularly suited to these type of systems [79, Sect. 10.5]. Considering the range of time-scale separations at which we obtain an acceptable accuracy, the results show that AMCA outperforms the other methods.

The rest of the chapter is structured as follows. Section 5.2 presents the

motivation and related work in this matter. Section 5.3 details the characteristics of the test networks analyzed and their associated CTMC models. Section 5.4 describes the quasi-stationary approximation and the related approximation methods based on time-scale separation. In Section 5.5 we present our approximation method called AMCA. Section 5.6 shows the numerical evaluation and the results. Finally, Section 5.7 draws the conclusions.

5.2 Motivation and related work

The analysis of wireless networks based on time-scale separation has been proposed in recent studies [84–91]. In many of them, the so-called QSA has been shown to be accurate and computationally efficient [85, 88–90]. However, when the gap between time-scales shortens, the accuracy of the method deteriorates to a point in which the method is no longer useful from a practical perspective.

In [86] a GQSA has been proposed. It can adjust the accuracy with a parameter called radius (R). In a recent study [92] we showed that, in some network scenarios, the accuracy achieved with GQSA improves as R increases. However, in other scenarios increasing R reduces the accuracy. More importantly, it is difficult to predict the scenarios in which the accuracy can be improved by increasing R.

In the proposed method AMCA, the Markov model of the network is structured in levels and phases. Then, we analyze the transient regime at each level to determine the fraction of time that the system spends at each of its phases until a level change occurs. Once these fractions of time are found for all phases in all levels, a new approximation of the stationary distribution of the complete system is computed. We repeat the procedure until a predefined accuracy is satisfied. This iterative procedure is initialized with the solution obtained by QSA.

To evaluate AMCA, we selected two test networks to apply the new approximation method to the same scenarios employed by previous approxi-

mate methods based on time-scale separation so that a fair comparison is carried out. Specifically, the CRN scenario was employed in [85] and the ISN scenario in [86]. However, the selection of these test networks does not limit the applicability of AMCA in any way.

5.3 Wireless networks description and modeling

In this section, we detail the main characteristics of the test networks. We describe the performance metrics of interest and define a two-dimensional CTMC model for each network.

5.3.1 Cognitive radio network

As in [85, 93], we model the primary user (PU) and secondary user (SU) traffic at the session (connection) level and ignore interactions at the packet level (scheduling, buffer management, etc.). We assume an ideal medium access control (MAC) layer for SUs, which allows a perfect sharing of the allocated channels among the active SUs (all active SUs get the same bandwidth portion), introduce zero delay and whose control mechanisms consume zero resources. In addition, we also assume that an active SU can sense the arrival of a PU in the same channel instantaneously and reliably. In this sense, the performance parameters obtained can be considered as an upper bound.

The cognitive radio network has C_1 *primary channels (PCs)* that can be shared by PUs and SUs, and C_2 *secondary channels (SCs)* only for SUs. Let $C = C_1 + C_2$ be the total number of channels in the network. Note that the SCs can be obtained from e.g., unlicensed bands, as proposed in [93]. This assumption is applicable to the *coexistence* deployment scenario for CRNs. Alternatively, as it might be of commercial interest for the primary and secondary networks to *cooperate*, the secondary channels may be obtained based on an agreement with the primary network [94]. A SU in the PCs might be forced to vacate its channel if a PU claims it to initiate a new session. As SUs support *spectrum*

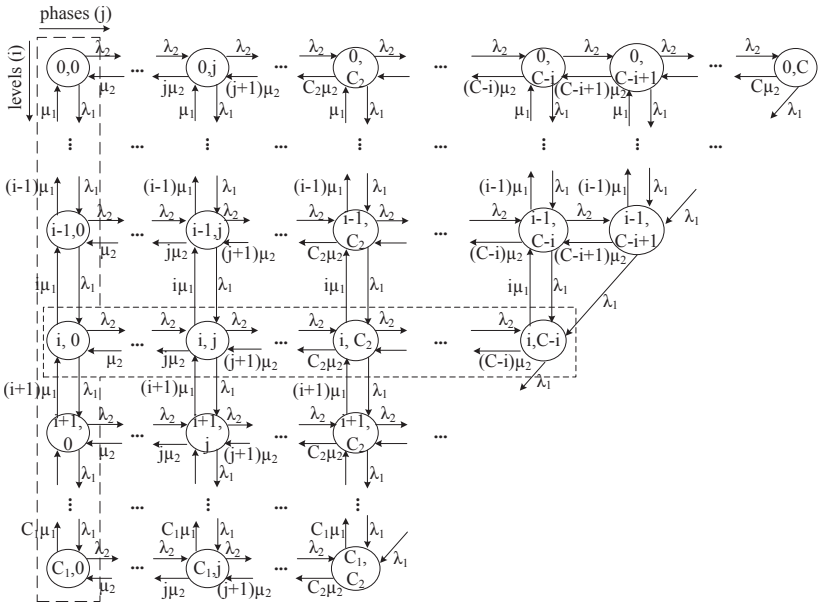


Figure 5.1: State-transition diagram, Cognitive Radio Network.

handover, a vacated SU can continue with its ongoing communication if a free channel is available. Otherwise, it is *forced to terminate*.

For the sake of mathematical tractability, Poisson arrivals and exponentially distributed service times are assumed. The arrival rate for PU (SU) sessions is λ_1 (λ_2), their service rate is μ_1 (μ_2), and requests consume 1 (1) channel when are accepted.

We denote by (i, j) the network state, when there are i ongoing PU sessions and j SU sessions. The set of feasible states is $\mathcal{S} \triangleq \{(i, j) : 0 \leq i \leq C_1, 0 \leq i + j \leq C\}$ and the cardinality of $\mathcal{S}$ is $|\mathcal{S}| = (C_1/2 + C_2 + 1)(C_1 + 1)$. The state-transition diagram of the network is depicted in Fig. 5.1. Given the set of feasible states and the transition rates among them, the global balance equations can be defined. Finally, the global balance equations together with the normalization equation can be solved to obtain the steady-state probabil-

ities denoted as $\pi(i,j)$.

The network performance parameters are determined as follows:

$$P_{pu} = \sum_{k=0}^{C_2} \pi(C_1,k) \quad , \quad P_{su} = \sum_{k=C_2}^{C} \pi(C-k,k), \tag{5.1}$$

$$P_{ft} = \frac{\lambda_1(P_{su} - \pi(C_1,C_2))}{\lambda_2(1-P_{su})}, \tag{5.2}$$

$$Th_{su} = \sum_{j=1}^{C} \sum_{i=0}^{Z} j\mu_2 \cdot \pi(i,j), \tag{5.3}$$

where P_{pu} is the PUs blocking probability, which clearly coincides with the one obtained in an Erlang-B loss model with C_1 servers; P_{su} is the SUs blocking probability, i.e., the fraction of SU sessions rejected upon arrival as they find the network full; P_{ft} is the forced termination probability of the SUs, i.e., the rate of SU sessions forced to terminate divided by the rate of accepted SU sessions; Th_{su} is the SUs throughput, i.e the rate of SU sessions successfully completed and $Z = \min(C_1, C-j)$.

5.3.2 Integrated service network

We use the same model defined in [83,86] for an integrated service network, where a single base station serves real-time (RT) and non-real-time (NRT) traffic. We consider that a link with a total capacity of C Mbps is shared among RT and NRT communications.

We assume that all RT calls (sessions) are of the same class and are given strict priority over the NRT traffic. We denote by N_{rt} the maximum number of channels for RT calls. When a RT call arrives, it occupies 1 channel (if available) of rate c bps. Note that a RT call occupies 1 channel during its entire service duration to meet its required QoS; otherwise, it is blocked. We set N_{rt}, such that $N_{rt} \cdot c$ is sufficiently smaller than C to avoid starvation of the NRT traffic. Let $n_{rt}(t)$ be the stochastic process number of RT calls in the network at time t, $t \geq 0$.

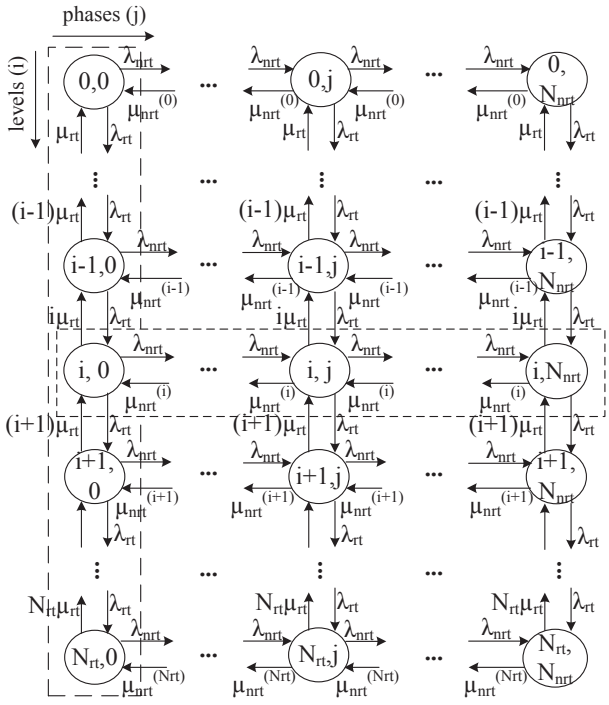


Figure 5.2: State-transition diagram, Integrated Service Network.

The capacity not used by the RT traffic is evenly shared by the NRT flows according to the processor sharing (PS) discipline. Let $n_{nrt}(t)$ be the stochastic process number of NRT flows in the network at time t, $t \geq 0$. Then, $\{(n_{rt}(t), n_{nrt}(t))\}$ is the joint RT and NRT stochastic process. The available capacity for the NRT traffic at time t is given by $C_{nrt}(t) = C - n_{rt}(t) \cdot c$. The bit-rate of each admitted NRT flow at time t is $c_{nrt}(t) = C_{nrt}(t)/n_{nrt}(t)$, and it is updated after any RT or NRT accepted arrivals or departures. To satisfy the QoS of admitted NRT flows, the maximum number of concurrent NRT flows is limited to N_{nrt}. Accordingly, an NRT flow arriving at time t is blocked if $n_{nrt}(t) = N_{nrt}$.For the sake of mathematical tractability, we assume Poisson arrivals for RT and NRT requests with rates λ_{rt} and λ_{nrt} respectively. Also,

the service time of each admitted RT request is exponentially distributed with rate μ_{rt}. The size of the flows generated by the NRT sessions are exponentially distributed with mean L (bits). Note that, the service time of NRT flows (transfer delay) depends on the available resources.

We denote by (i, j) the network state, when there are i ongoing RT calls and j NRT flows. Let $\mathcal{S}$ be the set of feasible states as $\mathcal{S} := \{(i, j) : 0 \leq i \leq N_{rt}, 0 \leq i + j \leq N_{rt} + N_{nrt}\}$ and the cardinality of $\mathcal{S}$ is $|\mathcal{S}| = (N_{rt} + 1)(N_{nrt} + 1)$. The state-transition diagram of the network is depicted in Fig. 5.2.

As before, given the set of feasible states and the transition rates among them, the global balance equations can be defined. Finally, the global balance equations together with the normalization equation can be solved to obtain the steady-state probabilities denoted as $\pi(i, j)$. Clearly, the service rate of NRT flows varies according to the number of RT calls in the network as:

$$\mu_{nrt}^{(i)} = \frac{C - i \cdot c}{L}. \tag{5.4}$$

The network performance parameters are determined by

$$P_{nrt} = \sum_{k=0}^{N_{rt}} \pi(k, N_{nrt}), \tag{5.5}$$

$$\mathbb{E}\left[X_{nrt}\right] = \sum_{j=1}^{N_{nrt}} \sum_{i=0}^{N_{rt}} j \cdot \pi(i, j), \tag{5.6}$$

$$\mathbb{E}\left[D_{nrt}\right] = \frac{E\left[X_{nrt}\right]}{\lambda_{nrt}(1 - P_{nrt})}, \tag{5.7}$$

where P_{nrt} is the blocking probability of NRT flows, $\mathbb{E}\left[X_{nrt}\right]$ is the mean number of NRT flows in the network and $\mathbb{E}\left[D_{nrt}\right]$ is the average transfer delay of NRT flows. Note that (5.7) is a direct application of Little's law.

5.4 Approximate solution methods

In this section, we describe the approximation methods based on time-scale separation that have appeared in the literature.

If the wireless network model in its entirety is too large or complex to analyze, the state-space may be partitioned into disjoint subsets of states. This partition is made by considering the essence of time-scale separation: the interactions among the states of a subset are strong (high frequency of events) but the interactions among the states of different subsets are weak (low frequency of events). Such models are sometimes referred to as nearly completely decomposable (NCD), nearly uncoupled, or nearly separable [79, 95].

Then, we organized the state-space of the CTMC of each test network into levels and phases (as shown in Fig. 5.1 for CRN and in Fig. 5.2 for ISN), so that we call levels the subsets in the y-axis and we call phases the states in the x-axis contained in the same level. The state transitions between states of the same level (phases) very often occur at a higher rate than the transitions between states of different levels, i.e., a high number of phase changes (in the same level) occur before a level change.

Next (Sections 5.4.1, 5.4.2, and 5.4.3) we detail how to compute the approximate steady-state probabilities with each of the approximation methods. With these approximated values, the performance parameters of each test network can be computed using (5.1)–(5.3) for the CRN, and (5.5)–(5.7) for the ISN.

5.4.1 Quasi-stationary approximation

The simplest approximation based on time-scale separation is the so called quasi-stationary (or, quasi-static) approximation (QSA) [84,88,89,96,97]. This approximation produces easily computable and accurate results when the separation of the time-scales is large.

We start by obtaining the probability distribution of finding the system at each level, i.e., the slow transitions (PUs in the CRN or RT traffic in the ISN) and denote it by

$$\boldsymbol{\pi} = [\pi(0) \;\; \pi(1) \;\; \cdots \;\; \pi(i) \;\; \cdots \;\; \pi(y)], \tag{5.8}$$

where y represents the highest level of the CTMC. Then, for each level, we proceed to obtain the conditional probability distributions of finding the system at each phase. This conditional distribution for level i is given as

$$\widehat{\boldsymbol{\pi}}(i) = [\widehat{\pi}(0|i) \;\; \widehat{\pi}(1|i) \;\; \cdots \;\; \widehat{\pi}(j|i) \;\; \cdots \;\; \widehat{\pi}(x|i)], \tag{5.9}$$

where x represents the highest phase in level i. These are approximate probability distributions because they are computed assuming that when the process enters a level, the time spent there is sufficiently large so that the stationary regime is reached.

Finally, the approximate stationary distribution of the system is computed using (5.8) and (5.9) as follows

$$\pi(i,j) \approx \widehat{\pi}(i,j) = \pi(i) \cdot \widehat{\pi}(j|i). \tag{5.10}$$

5.4.2 Generalized quasi-stationary approximation

In GQSA [86], the system stationary distribution can be approximated as in QSA, but now a set of adjacent levels is considered for the analysis of level i, rather than just level i. For that, the parameter R indicates the number of adjacent levels to consider. Clearly, R allows to adjust the trade-off between accuracy and computational cost.

The number of levels required at each GQSA step is equal to $2R+1$. Note that QSA can be seen as a special case of GQSA with $R = 0$.

Let $\Omega(i)$ be the set of states contained in level i and its $2R$ closest levels and denote by $\pi_{\Omega(i)}(i,j)$ the stationary distribution of the CTMC restricted to the states in $\Omega(i)$ and the transitions between them. Then, the approximate stationary distribution of the system (i,j) is computed as follows

$$\pi(i,j) \approx \overline{\pi}(i,j) = \pi(i) \cdot \frac{\pi_{\Omega(i)}(i,j)}{\sum_j \pi_{\Omega(i)}(i,j)}. \tag{5.11}$$

5.4.3 Iterative aggregation/disaggregation approximation

In the IAD method, as with QSA, the idea is to partition the state-space into aggregates (subsets of states), estimate the probability that the system is in a particular aggregate, estimate the conditional probabilities of being in each state of every aggregate, and then combine them to obtain an approximation of the stationary distribution of the complete system [79, Chap.10], [98].

In our test networks, the transition rate matrix $\boldsymbol{Q}$ has the following NCD block structure:

$$\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{D}_1 & \boldsymbol{U}_1 & & & \\ \boldsymbol{L}_2 & \boldsymbol{D}_2 & \boldsymbol{U}_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \boldsymbol{L}_{n-1} & \boldsymbol{D}_{n-1} & \boldsymbol{U}_{n-1} \\ & & & \boldsymbol{L}_n & \boldsymbol{D}_n \end{bmatrix},$$

where $\boldsymbol{D}_n$, $\boldsymbol{U}_n$, and $\boldsymbol{L}_n$ are matrices of suitable dimension that aggregate the diagonal, upper, and lower transition rates of level n, respectively. Please refer to [79, Chap.10] for specific details on the matrix $\boldsymbol{Q}$ NCD block structure. Next we detail our specific implementation of the IAD method:

1. Use the QSA to determine the initial stationary distribution $\boldsymbol{\pi}_i^{(0)}$.

2. Apply the following iteration until the convergence test is met:

$$\boldsymbol{\pi}_i^{(k+1)} = \begin{cases} \boldsymbol{\pi}_{i+1}^{(k)} \boldsymbol{W}_i, & i = 1, \\ \boldsymbol{\pi}_{i-1}^{(k+1)} \boldsymbol{V}_i + \boldsymbol{\pi}_{i+1}^{(k)} \boldsymbol{W}_i, & i = 2, \ldots, n-1 \\ \boldsymbol{\pi}_{i-1}^{(k+1)} \boldsymbol{V}_i, & i = n, \end{cases} \tag{5.12}$$

where $\boldsymbol{\pi} = [\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_n]$, $\boldsymbol{\pi}_i$ is the sub-vector of $\boldsymbol{\pi}$ that corresponds to level i, $\boldsymbol{\pi}_i^{(k)}$ is the sub-vector value at the k-th iteration, $\boldsymbol{V}_i = -U_{i-1} D_i^{-1}$ and $\boldsymbol{W}_i = -L_{i+1} D_i^{-1}$.

Convergence test

Using the solution obtained by the QSA as $\boldsymbol{\pi}^{(0)}$, the iterative procedure terminates when the following convergence test is met:

$$\hat{e}_r(z^{(k)}) \;=\; \frac{|z^{(k-1)} - z^{(k)}|}{z^{(k)}} \leq \varepsilon\,, \tag{5.13}$$

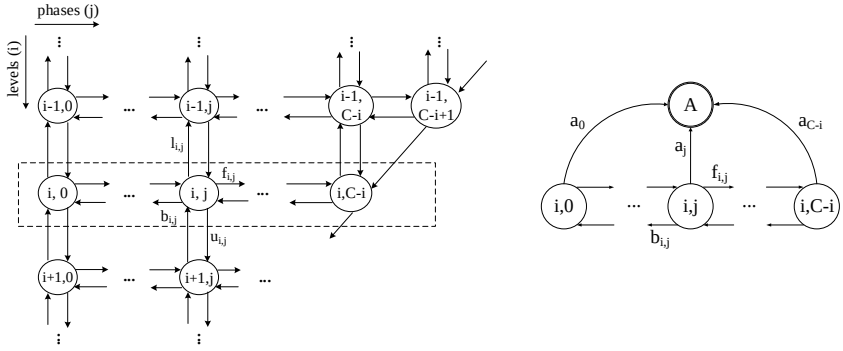
where z is one of the performance metrics in $\{P_{su}, P_{ft}, Th_{su}\}$ for the CRN evaluation, or in $\{P_{nrt}, \mathbb{E}[D_{nrt}]\}$ for the ISN evaluation. We iterate until the estimated error for *all* performance parameters of the test network is less than a predefined ε.

5.5 Absorbing Markov chain approximation

In this section, we present the proposed AMCA, which is an iterative method. As in previous methods, in AMCA we also structure the CTMCs of the test networks in levels and phases. We analyze the system in the transient regime and model the time spent by the system in a level as a phase-type distribution. For each level, we determine the fraction of time the system spends in each of the phases of the level, i.e., since entering the level until departing from it. For doing so, AMCA requires to know the probabilities of finding the system in each phase of the adjacent levels. Thus, an iterative method is devised, that is terminated using the same convergence test detailed in 5.4.3. Next, we detail the procedure, the equations and variables involved in our method, and finally present the AMCA algorithm.

5.5.1 Approximation method

In the QSA it is assumed that, when the process enters a level, it takes an infinitely long time to leave that level. In our method, we assume that although the sojourn time in a level will be typically large (consistently with the large separation between time-scales), it is finite. Then, we obtain the probability



(a) CRN, level i and its adjacent levels

(b) CRN, absorbing Markov chain of level i with the absorption state A

Figure 5.3: Transitions and absorption state.

that the process is in phase j of level i, given that the process is in level i, as the fraction of time that the process spends in phase j during a sojourn of the CTMC in level i.

In order to study the sojourn time in a level, we model the states of each level i of the original CTMC as transient states of an absorbing Markov chain with an absorbing state A, where all states outside level i have been lumped together. To illustrate this, in Fig. 5.3a we represent a region of the CRN state-transition diagram (level i and its adjacent levels) and in Fig. 5.3b the absorbing Markov chain corresponding to level i. As a result, the outgoing transitions from a state (i, j) that in the original CTMC go to a state inside level i (those with rates $b_{i,j}$ and $f_{i,j}$) are directly mapped onto the absorbing Markov chain. In contrast, all transitions from a state (i, j) that in the original CTMC go to a state outside level i (those with rates $l_{i,j}$ and $u_{i,j}$) are aggregated into a single transition in the absorbing Markov chain that leads to the absorbing state (i.e., $a_j = l_{i,j} + u_{i,j}$).

Note that, if we knew the probabilities with which the sojourn time in a level is initiated at each of its phases, then the conditional probabilities

obtained by this method would be the exact ones. However, unless the original CTMC has some special structure (for instance, if each level can only be entered by exactly one of its phases), these initial probabilities cannot be obtained without having the stationary distribution of the whole CTMC.

We propose to use the QSA to estimate the initial stationary distribution of the complete system. Then, we obtain the fractions of time spent at each of the phases of each level before absorption. Finally, we combine the estimation of the conditional probabilities of finding the system at the phases of each level, and the probability distribution of finding the system at each level, to determine a new approximation for the stationary distribution of complete system. This way, we obtain a refinement of the initial approximate stationary distribution. The same process can be repeated iteratively to further improve the approximation.

Based on the basic properties of PH distributions (see B.3), the iterative procedure described above is defined by the following equations:

$$\boldsymbol{v}_i^{(k-1)} = \pi_{i-1}(\widetilde{\boldsymbol{\pi}}_{i-1}^{(k-1)}\boldsymbol{U}_{i-1}) + \pi_{i+1}(\widetilde{\boldsymbol{\pi}}_{i+1}^{(k-1)}\boldsymbol{L}_{i+1}) \,, \tag{5.14}$$

$$\boldsymbol{\alpha}_i^{(k-1)} = \left[\boldsymbol{v}_i^{(k-1)}\boldsymbol{e}\right]^{-1}\boldsymbol{v}_i^{(k-1)} \,, \tag{5.15}$$

$$\widetilde{\boldsymbol{\pi}}_i^{(k)} = \left[\boldsymbol{\alpha}_i^{(k-1)}(-T_i^{-1})\boldsymbol{e}\right]^{-1}\boldsymbol{\alpha}_i^{(k-1)}(-T_i^{-1}) \,, \tag{5.16}$$

where

- the superscript (k) denotes the iteration number and $\boldsymbol{e}$ is a column vector of ones of appropriate dimension.
- $\boldsymbol{v}_i^{(k-1)}$ is a row vector that contains the input rates to each state of the level i. Its initial value is given by

$$\boldsymbol{v}_i^{(0)} = \pi_{i-1}(\widetilde{\boldsymbol{\pi}}_{i-1}^{(0)}\boldsymbol{U}_{i-1}) + \pi_{i+1}(\widetilde{\boldsymbol{\pi}}_{i+1}^{(0)}\boldsymbol{L}_{i+1}), \tag{5.17}$$

  where $\boldsymbol{U}_{i-1}$ is a matrix of suitable dimension with the transition rates from level $i-1$ to level i and $\boldsymbol{L}_{i+1}$ is a matrix of suitable dimension with the transition rates from level $i+1$ to level i.

- $\boldsymbol{\alpha}_i^{(k-1)}$ is the initial probability row vector for level i, i.e., the j-th element of this vector, $\alpha_i^{(k-1)}(j)$, is the probability that the process enters through phase j when it visits level i. Its initial value is given by

$$\boldsymbol{\alpha}_i^{(0)} = \left[\boldsymbol{v}_i^{(0)}\boldsymbol{e}\right]^{-1}\boldsymbol{v}_i^{(0)}. \tag{5.18}$$

- $\widetilde{\boldsymbol{\pi}}_i^{(k)}$ is a row vector containing the fractions of time the process spends in each phase of level i before absorption, e.g., the j-th element of this vector, $\widetilde{\pi}_i^{(k)}(j)$, is the fraction of time the process spends in the phase j of level i before absorption. Its initial value is given by QSA

$$\widetilde{\boldsymbol{\pi}}_i^{(0)} = \widehat{\boldsymbol{\pi}}(j|i), \tag{5.19}$$

  where $\widehat{\boldsymbol{\pi}}(j|i)$ is the distribution of probabilities of

  - CRN: finding j ongoing SU sessions in an M/M/(C$-i$)/(C$-i$) system with only SUs.
  - ISN: finding j NRT flows in an M/M/1/N-PS system with only NRT traffic.

- π_i is the probability of finding the system at level i. It is the probability of finding i PUs in the CRN or i ongoing RT sessions in the ISN. It is computed using simple recursions since their corresponding CTMC are one-dimensional birth-and-death processes.

Finally, the steady-state probability distribution can be approximated as

$$\pi(i,j) \approx \widetilde{\pi}^{(k)}(i,j) = \pi_i \cdot \widetilde{\pi}_i^{(k)}(j)\,. \tag{5.20}$$

To compute the approximate values of the performance parameters, we use (5.1)–(5.3) for the CRN, and (5.5)–(5.7) for the ISN, with the distribution of probabilities defined in (5.20). Finally, the proposed iterative method may be halted once the predefined convergence test defined in Section 5.4.3 is satisfied.

Algorithm 2 summarizes the procedure used to conduct the performance evaluation of the test networks with AMCA.

Algorithm 2: Iterative Absorbing Markov Chains Approximation Method

1 Set $\pi^{(0)} \approx \widehat{\pi}(i,j) = \pi(i) \cdot \widehat{\pi}(j|i)$ as the initial approximation to the $\boldsymbol{\pi}$ solution computed by QSA. Set $k = 1$.
2 Compute the vector of input rates to each state of the level i, $\boldsymbol{v}_i^{(k-1)}$; (5.14)
3 Compute the initial probability vector of level i, $\boldsymbol{\alpha}_i^{(k-1)}$; (5.15)
4 Compute the conditional probabilities vector, $\widetilde{\boldsymbol{\pi}}_i^{(k)}$; (5.16)
5 Compute the new approximate steady-state probability distribution $\pi(i,j) \approx \widetilde{\pi}^{(k)}(i,j)$; (5.20)
6 Compute the performance metrics of each network: using (5.1)–(5.3) for CRN, and (5.5)–(5.7) for ISN;
7 Apply the convergence test to iterations k and $k-1$: (5.13)
8 **if** *satisfactory* **then**
9 | stop
10 **else**
11 | set $k = k + 1$;
12 | go to step 2;
13 **end**

5.6 Numerical evaluation and results

We perform two types of analysis. First, we evaluate the behavior of the approximation methods when the separation of time-scales varies. Second, we study the trade-off between accuracy and computational cost; the results of these analyses are presented in Sections 5.6.1 and 5.6.2, respectively.

As a baseline for our study, we implemented the exact solution of the CTMC associated with each test network, in order to evaluate the error of the approximation methods. For the sake of comparison, we used test network sizes that allowed the computation of the exact solution with reasonable execution time and memory requirements. In addition, we implemented GQSA and the IAD method to validate AMCA and to compare its performance with that of the other methods in terms of accuracy and computational cost.

The accuracy of the methods is measured as the relative error (e_r) of each performance parameter. For instance, the relative error of the SUs blocking probability in the CRN is computed as

$$e_r(P_{su}) = \frac{|P_{su}^E - P_{su}^A|}{P_{su}^E}, \tag{5.21}$$

where P_{su}^E is the exact SUs blocking probability and P_{su}^A is the approximate SUs blocking probability. Note that (5.21) is the (exact) relative error whereas $\hat{e}_r(P_{su}^{(k)})$, as defined in (5.13), is an estimation of it.

We evaluate the performance of the test networks for different sizes (number of channels available for each type of user or flow) and different load conditions. For the SUs in the CRN, we analyze their blocking probability, forced termination probability and throughput. We consider the following values for the number of primary channels: $C_1 \in \{70, 80, 90, 100, 120, 140\}$. For each of them, we consider the following values for the number of secondary channels: $C_2 \in \{10, 20, 40, 60\}$.

For the NRT traffic in the ISN, we determine its blocking probability and the average transfer delay. Keeping $c = 64\,\text{kb/s}$ and $L = 500\,\text{kB}$ constant, we consider the following values for the total link capacity of the network: $C \in \{1.92, 7.68, 10\}$ Mbps, which are a similar to the ones used in [86].

We set the service rates to $1\,\text{s}^{-1}$, and then we adjust the arrival rates to obtain two load conditions: low (L) and high (H), which correspond to blocking probabilities of $1 \cdot 10^{-3}$ and $5 \cdot 10^{-2}$, respectively. Combining the two load conditions for each user type or traffic category, we obtain four different configurations:

LL low load condition for PUs (RT traffic), and low load condition for SUs (NRT traffic).

LH low load condition for PUs (RT traffic), and high load condition for SUs (NRT traffic).

HL high load condition for PUs (RT traffic), and low load condition for SUs (NRT traffic).

HH high load condition for PUs (RT traffic), and high load condition for SUs (NRT traffic).

In the following, we present the results obtained for the two types of analysis.

5.6.1 Behavior of the approximation methods when the separation of time scales varies

We analyze the behavior of the approximation methods as a function of the time-scale separation. For that, we first configure the test networks to a specific load condition (LL, LH, HL or HH). Then, we use an accelerating factor f, $10^{-5} \leq f \leq 10^{5}$, to equally accelerate or decelerate both the arrival and service rates of the components with high priority in the networks (PUs in the CRN or RT traffic in the ISN), while keeping the offered traffic constant. For instance, in the CRN, for each value of f the PU arrival and service rates are obtained as $\lambda_1(f) = f\,\lambda_1$ and $\mu_1(f) = f\,\mu_1$. Note that the offered traffic $\lambda_1(f)/\mu_1(f) = \lambda_1/\mu_1$ is independent of f. As f approaches 0 the event rate of high priority users gets lower. Therefore, the behavior of the systems gets better aligned with the hypothesis underlying all approximation methods considered here: high priority users are nearly static from the perspective of low priority users. As a consequence, it is expected that the accuracy of all approximation methods improves when f decreases toward 0, and conversely, degrades when f grows.

In Figs. 5.4–5.7 we show the relative error of the blocking probability against the accelerating factor f for LH and LL load conditions. With regard to the other performance metrics and load conditions, the behavior of the approximation methods is qualitatively similar, but for conciseness their results are not shown.

We can quantify the validity range of an approximation in the time-scale domain as the maximum value of f for which a certain accuracy is met. Figures 5.4 and 5.5 show that AMCA can extend the validity range of QSA at the expense of higher computational cost.

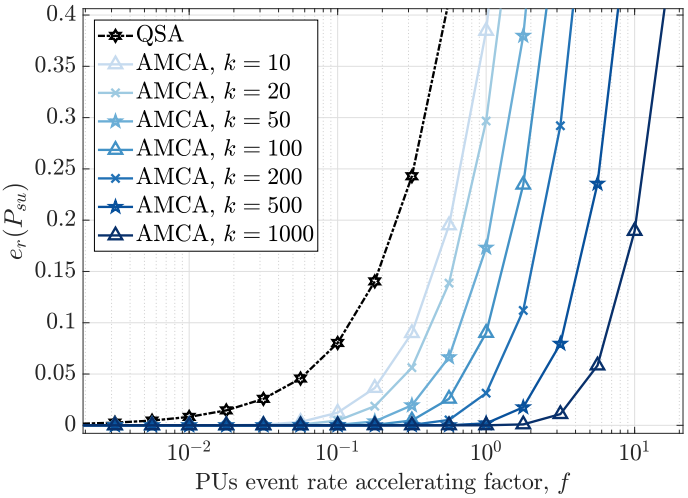


Figure 5.4: Relative error for the SUs blocking probability in LH load condition: $\lambda_1 = 110.90\,\mathrm{s}^{-1}$, $\mu_1 = 1\,\mathrm{s}^{-1}$, $C_1 = 140$; $\lambda_2 = 87.69\,\mathrm{s}^{-1}$, $\mu_2 = 1\,\mathrm{s}^{-1}$, $C_2 = 60$; k denotes the number of iterations performed.

In Figs. 5.6 and 5.7 we compare AMCA with GQSA and IAD in terms of accuracy at different time-scales. These results were obtained by the following procedure. We measured the time to execute GQSA with a given radius R ($\mathrm{GQSA_R}$). Then, for IAD and AMCA we performed the maximum number of iterations such that computation time not higher than computation time of $\mathrm{GQSA_R}$. These results are labeled as $\mathrm{AMCA_R}$ and $\mathrm{IAD_R}$. For instance, the curve for $\mathrm{AMCA_1}$ represents the result obtained iterating AMCA while the computation time not exceeding that of $\mathrm{GQSA_1}$.

The following observations can be made from Figs. 5.6 and 5.7:

- As expected, with all approximation methods, when the accelerating factor f decreases ($f \to 0$) the approximate values of all evaluated performance parameters tend to their exact values.

- Increasing the radius in GQSA not always ensures a reduction of the

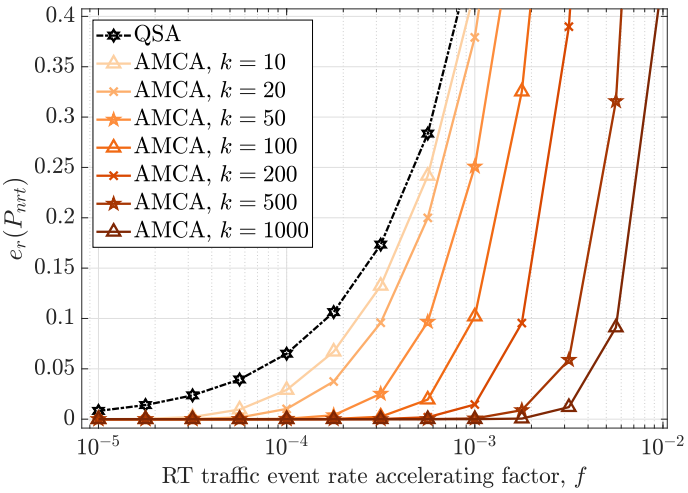


Figure 5.5: Relative error for the NRT flows blocking probability in LL load condition: $\lambda_{rt} = 75.24\,\mathrm{s}^{-1}$, $\mu_{rt} = 1\,\mathrm{s}^{-1}$, $N_{rt} = 100$; $\lambda_{nrt} = 1.27\,\mathrm{s}^{-1}$, $N_{nrt} = 140$; $C = 10$ Mbps, $c = 64$ kbps, $L = 4$ Mb; k denotes the number of iterations performed.

relative error [92]. Figure 5.6 illustrates this behavior; as can be seen for $f > 10^{-1}$, GQSA_1 has better accuracy than GQSA_2 and GQSA_3.

- AMCA outperforms GQSA and IAD in terms of validity range, i.e., with the same computation time AMCA is able to achieve a validity range wider than that of GQSA and IAD method. For instance, see in Fig.5.7 the curves of QSA, GQSA_3, IAD_3 and AMCA_3; for a relative error lower than 0.05, AMCA is able to achieve a validity range that is approximately 18 times wider than that of QSA, whereas GQSA and IAD method are able to achieve a validity range of approximately 2 and 6 times wider than that of QSA, respectively. A similar behavior was observed for all load conditions and network sizes.

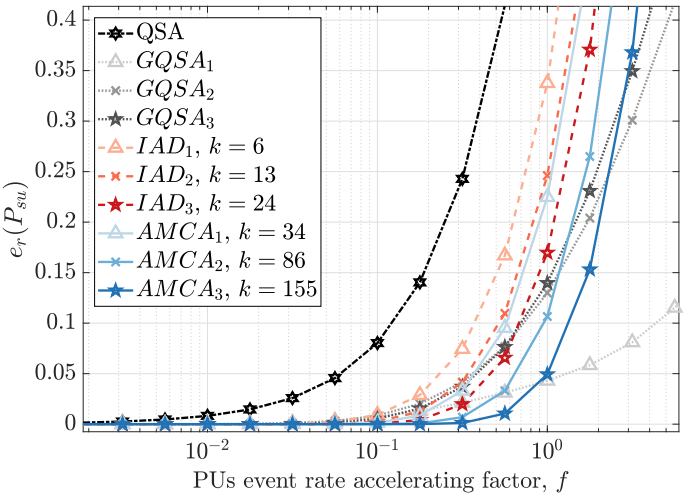


Figure 5.6: Relative error for the SUs blocking probability in LH load condition: $\lambda_1 = 110.90\,\mathrm{s}^{-1}$, $\mu_1 = 1\,\mathrm{s}^{-1}$, $C_1 = 140$; $\lambda_2 = 87.69\,\mathrm{s}^{-1}$, $\mu_2 = 1\ \mathrm{s}^{-1}$, $C_2 = 60$; k denotes the number of iterations performed.

5.6.2 Trade-off between accuracy and computational cost

In this section, we analyze the trade-off between accuracy and computational cost. Figures 5.8 (CRN) and 5.9 (ISN) illustrate the evolution of the relative error of the blocking probability with the execution time. To obtain these results, we set f such that the relative error obtained by QSA for the studied parameter is 10%. Recall that the stationary distribution obtained by QSA is used for the initial values of IAD and AMCA.

It is worth nothing that GQSA is not an iterative method in the sense that it can be executed for any radius value, $R = n$, without having previously obtained the results for $R = 0, 1, \ldots, n-1$. However, there is no available method that allows to find the appropriate n to achieve a given accuracy. Although it does not always occur (for example in Fig. 5.8 the accuracy decreases from GQSA_1 to GQSA_2), it is expected that the obtained accuracy tend to improve when R is increased. Thus, in our comparative study we in-

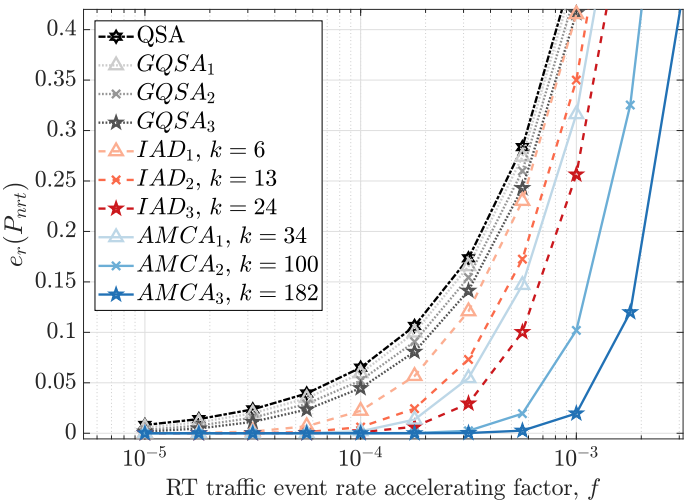


Figure 5.7: Relative error for the NRT flows blocking probability in LL load condition: $\lambda_{rt} = 75.24\,\mathrm{s}^{-1}$, $\mu_{rt} = 1\,\mathrm{s}^{-1}$, $N_{rt} = 100$; $\lambda_{nrt} = 1.27\,\mathrm{s}^{-1}$, $N_{nrt} = 140$; $C = 10$ Mbps, $c = 64$ kbps, $L = 4$ Mb; k denotes the number of iterations performed.

crease R until a predefined convergence test (estimated relative error) is met, roughly mimicking the operation of the other two iterative methods.

We can observe in Figs. 5.8 and 5.9 how the accuracy of each method evolves as the number of iterations, and hence the computation time, increases. Note that the execution time to obtain a determined accuracy with AMCA is lower than that of GQSA and IAD.

Tables 5.1 and 5.2 show the relative error estimations ($\hat{e}_r$), the execution times of all the approximation methods, and (for comparative effects) the required time to obtain the exact system stationary distribution for the CRN and the ISN test networks, respectively. To obtain these results, we have considered scenarios where the initial (exact) relative errors obtained by QSA were 20% and 40%, and with different load configurations. These two values represent scenarios in which the separation between time-scales is not long

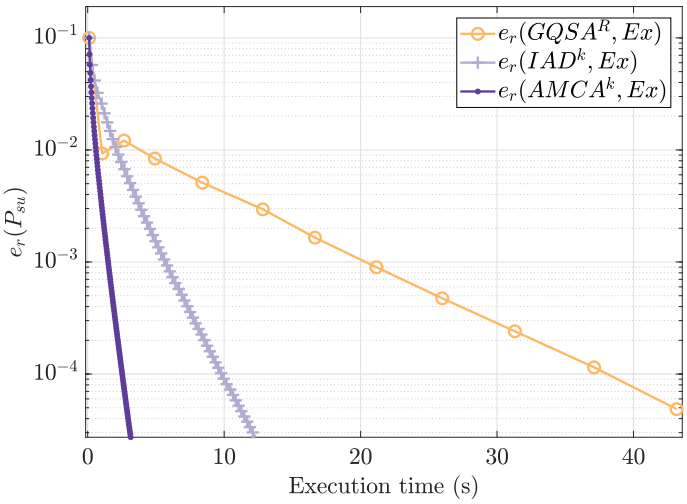


Figure 5.8: Relative error for the SUs blocking probability in LH load condition: $\lambda_1 = 13.90\,\mathrm{s}^{-1}$, $\mu_1 = 0.13\,\mathrm{s}^{-1}$, $C_1 = 140$; $\lambda_2 = 87.69\,\mathrm{s}^{-1}$, $\mu_2 = 1\,\mathrm{s}^{-1}$, $C_2 = 60$.

Table 5.1: Relative Error Analysis - Cognitive Radio Network

Load Config.	$e_r^{(0)}$ (%)	$\hat{e}_r^{(k)}$			Execution Time (s)			
		GQSA	IAD	AMCA	GQSA	IAD	AMCA	Exact
LL	20	9.1e-4	9.9e-6	9.9e-6	184.9	134.5	10.4	209.1
LL	40	0.01	2.1e-5	9.9e-6	197.9	209.8	43.7	210.6
LH	20	6.6e-4	9.9e-6	9.9e-6	88.1	25.5	3.9	207.5
LH	40	9.8e-4	9.9e-6	9.9e-6	14.3	55.7	8.1	208.8
HL	20	7.1e-4	9.9e-6	3.9e-6	91.2	79.0	0.8	207.3
HL	40	6.3e-4	9.9e-6	9.9e-6	132.5	187.1	55.0	209.5
HH	20	6.8e-4	9.6e-6	9.8e-6	54.8	30.3	4.4	207.5
HH	40	7.0e-4	9.9e-6	9.7e-6	100.4	65.0	9.2	212.1

enough so that QSA cannot provide a sufficiently accurate approximation and, as a consequence, an enhanced method is required.

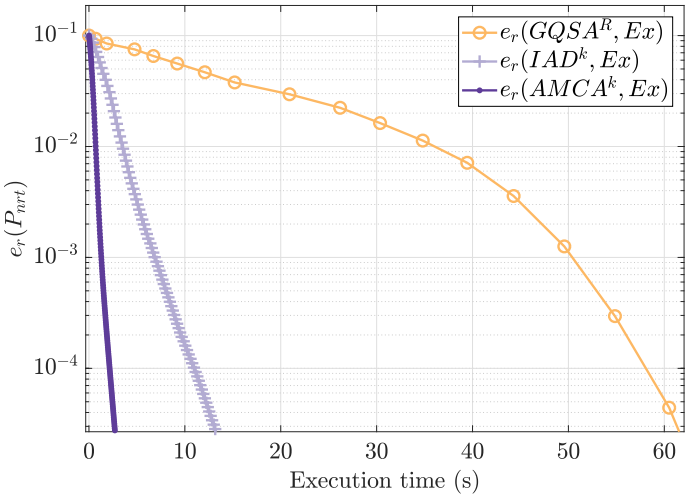


Figure 5.9: Relative error for the NRT flows blocking probability in LL load condition: $\lambda_{rt} = 1.25 \cdot 10^{-2}\,\mathrm{s}^{-1}$, $\mu_{rt} = 1.65 \cdot 10^{-4}\,\mathrm{s}^{-1}$, $N_{rt} = 100$; $\lambda_{nrt} = 1.27\,\mathrm{s}^{-1}$, $N_{nrt} = 140$; $C = 10$ Mbps, $c = 64$ kbps, $L = 4$ Mb.

Table 5.2: Relative Error Analysis - Integrated Service Network

Load Config.	$e_r^{(0)}$ (%)	$\hat{e}_r^{(k)}$			Execution Time (s)			
		GQSA	IAD	AMCA	GQSA	IAD	AMCA	Exact
LL	20	9.6e-4	9.7e-6	9.9e-6	53.4	25.7	3.1	118.8
LL	40	7.4e-4	9.9e-6	9.9e-6	61.9	46.8	6.3	120.0
LH	20	3.5e-4	4.9e-6	9.5e-6	96.1	23.5	3.4	118.5
LH	40	6.4e-3	1.1e-4	9.9e-6	108.9	118.4	35.2	118.8
HL	20	7.7e-4	9.9e-6	4.3e-6	20.5	12.1	1.2	118.8
HL	40	5.8e-4	9.6e-6	8.8e-6	34.6	24.7	2.3	118.8
HH	20	9.7e-4	7.6e-6	5.0e-6	38.1	15.8	2.8	118.8
HH	40	3.6e-4	9.7e-6	9.5e-6	77.6	69.1	9.2	118.9

In practice, when the exact value of the error is not available, a stop criterion is needed for the normal use of the approximation methods. Here the

estimated relative error ($\hat{e}_r$) is used as the stop criterion.

The threshold that $\hat{e}_r$ must fall below for the procedure to stop is chosen heuristically such that: (i) the (exact) relative error obtained after the iterative procedure stops must be $e_r \leq 10^{-2}$; (ii) it is unique for each method and for all the studied configurations. Note however that $\hat{e}_r$ is not necessarily the same for all the methods; the $\hat{e}_r$ values used for GQSA, IAD, and AMCA are: 10^{-3}, 10^{-5}, and 10^{-5}, respectively. In addition, the iterative procedure is halted by time, i.e., we established a maximum execution time so that, the iterative procedure stops when the time required to meet the convergence test is larger than the maximum execution time.

We observe that AMCA converges in all the evaluated scenarios with significantly lower execution times than those of GQSA and IAD. In the CRN case, AMCA is between 2 and 114 times faster than GQSA, and between 3 and 99 times faster than IAD. A similar behavior was observed in the ISN case: AMCA is between 3 and 28 times faster than GQSA and between 3 and 11 times faster than IAD. Note that there are a couple of scenarios (see Table 5.1, row LL – 40% and Table 5.2, row LH – 40%) in which GQSA and IAD were halted by time. Although in such cases converge could have been achieved if more iterations had been performed, it would be of no practical interest, since the benefit with respect to the exact solution (in terms of execution time) will be marginal or non-existent.

5.7 Highlights

In this chapter, we have presented a novel approximation method named absorbing Markov chain approximation (AMCA) for the performance evaluation of wireless networks that is based on time-scale separation. AMCA, which is iterative in nature, permits trading-off computational effort in exchange of an increased accuracy. We applied AMCA to analyze two different types of wireless networks and we compared the performance of our method with that of a recently published generalized quasi-stationary approxima-

tion (GQSA) and also with a classical method known as iterative aggregation/disaggregation (IAD). Numerical results show that our method outperforms GQSA and IAD by providing the same accuracy with substantially lower computational cost.

Chapter 6

Conclusions and Future Perspectives

In this dissertation, we contributed to the congestion control in the radio access network and core network of wireless cellular systems and to the performance evaluation techniques of large multi-service networks. We studied the capacity of the random access channel and the impact of the configuration parameters of the network in the random access procedure. Besides, we studied in-depth the access class barring (ACB) mechanism and proposed algorithms for its dynamic and efficient deployment. These algorithms can be exploited in current and forthcoming cellular technologies, such as narrow band IoT (NB-IoT) and 5th generation (5G) New Radio. Our solutions address the challenging task of autonomously activating/deactivating the barring scheme and tuning its configuration parameters so that it can dynamically and optimally adapt the incoming traffic load to the network capacity. Our solutions conform with current system specifications and can be integrated into real network equipment allowing efficient congestion control, and facilitating the coexistence of human-to-human (H2H) and massive machine-to-machine (M2M) traffic. In addition, we introduce a unified software-defined platform for 5G systems, the so-called SoftAir, and software-defined gateways (SD-GWs) were designed aiming at: (i) exploring the interactions between two types of networks [i.e., Internet of Things (IoT) and software-defined RANs

(SD-RANs)], and (ii) satisfying the requirements of nowadays IoT applications, concerning low latency in data delivery, efficient energy consumption, high scalability to accommodate a large number of devices, and ubiquitous connectivity for users. Finally, we proposed a novel and computationally efficient solution method to evaluate the performance of multi-service networks with multiple types of traffic flows that evolve at different time-scales. This method is useful when analyzing continuous-time Markov chain (CTMC) network models with a large state-space as it considerably reduces the computational cost required to obtain common performance parameters. It also allows to understand the interactions among different components of these rather complex networks.

The contributions in each chapter are summarized as follows:

- In Chapter 2, we conduct a thorough performance analysis of both the LTE Advanced (LTE-A) random access channel and the ACB as defined in the 3rd Generation Partnership Project (3GPP) specifications. Specifically, we seek to enhance the performance of LTE-A in massive M2M scenarios by modifying certain configuration parameters and by the implementation of ACB. We observed that ACB is appropriate for handling sporadic periods of congestion. Concretely, our results reflect that the access success probability of M2M UEs in the most extreme test scenario suggested by the 3GPP improves from approximately 30%, without any congestion control scheme, to 100% by implementing ACB and setting its configuration parameters properly.

- In Chapter 3, we deal with optimizing the ACB mechanism and proposing novel algorithms for its dynamic operation since there is no specification regarding how to dynamically adapt the barring parameters to the offered traffic so that network overload is avoided and a better QoS can be provided. This is useful for the efficient provisioning of simultaneous H2H/M2M communications, especially in highly changing scenarios with bursty traffic, as it can occur when M2M communications are involved. Two approaches for its dynamic implementation are

proposed. First, a dynamic algorithm based on reinforcement learning is devised, it can adapt the barring rate to different traffic conditions, reducing congestion and hence the number of collisions in the RACH. Second, an approach in which an estimate of the current number of M2M devices in backoff state is used to adjust in real-time the barring rate parameter. We validate our algorithms by extensive simulations in several scenarios with different degrees of traffic load. Numerical results show that congestion episodes are well-managed when using our proposed solutions. The performance advantage of the proposed dynamic solutions is quantified using common key performance indicators (KPIs).

- In Chapter 4, the so-called SoftAir architecture on wireless software-defined networks is introduced, and SD-GWs that jointly optimize cross-layer communication functionality between heterogeneous IoT devices and cellular systems are proposed. First, the SoftAir architecture is proposed to support a unified software-defined platform for quality-of-service aware IoT systems and SD-RANs with millimeter-wave transmissions. Next, the SD-GWs are designed in SoftAir to explore the interactions between two types of networks (i.e., IoTs and SD-RANs) and enable cross-layer solutions that simultaneously achieve optimal energy savings and throughput gain in IoTs and maximum sum-rates in SD-RANs. Simulation results validate that our SoftAir solutions surpass classical IoT schemes by jointly optimizing communication functionality for both IoTs and SD-RANs and bring significant system synergies for reliable 5G IoT communication.

- In Chapter 5, a new approximate method named absorbing Markov chain approximation (AMCA) is developed for the analysis of Markov chains based on the separation of time scales that improves the performance of existing methods. This analysis technique is particularly useful for systems that give rise to models with a large number of states, and in which events occur at different time scales, such as in

multi-service networks. We model the time the system spends in a series of subsets of states by a phase-type distribution and, for each of them, determine the probabilities of finding the system in each state of this subset until absorption. We compare the AMCA performance to that obtained by classical methods and to a recently proposed approach that aims at generalizing the conventional quasi-stationary approximation. We find that AMCA has a more predictable behavior, applies to a broader range of time-scale separations, and achieves higher accuracy for a given computational cost.

Several future directions of research arise out of this dissertation. In particular, the open research lines include:

- Perform a comprehensive study of the random access procedure in new technologies such as NB-IoT or 5G New Radio so that we can deepen our understanding and extend the usability of our algorithms devised in Chapter 3. The aim would be to find optimal configurations for such technologies in new scenarios with different traffic models and application requirements.
- Perform an in-depth study of an additional congestion control scheme, namely extended access barring, that has been specified by the 3GPP for delay-tolerant M2M devices and compare its performance with that of ACB. The objective would be to identify the scenarios where one mechanism excels the other and the scenarios in which any of them is suitable.
- Enable and explore the emerging concept of network slicing in the architecture developed in Chapter 4 so that the needs of a wide range of applications, with very different requirements and in heterogeneous and very diverse scenarios can be satisfied. Network slicing would allow each tenant (i.e., operator, provider, or user) to view the slice (i.e., subset of network resources) as a separate logical network with properties similar to those of a dedicated network. This perception would be

maintained even when the slices are actually implemented on a common processing, transport, and radio infrastructure operated by an infrastructure provider.

- Explore resource management mechanisms in evolved network infrastructures of 5G deployments that support the concept of network slicing, and the application of automatic learning techniques to optimize network configuration, especially in dense and heterogeneous scenarios.
- Apply the proposed method in Chapter 5, AMCA, in the performance analysis of networks with different classes of M2M devices in which the access attempts of each class is performed at different time-scales, or in cases where the M2M and H2H communications share available network resources.

Appendices

Appendix A

Notations

$\mathbf{x}$	boldface lower case symbols represent vectors
$\mathbf{X}$	boldface upper case symbols represent matrices
$\mathbf{I}_x$	denote an x by x identity matrix
$\mathbb{C}^{x,y}$	denote the set of $x \times y$ complex matrices
$\mathrm{tr}(\cdot)$	denote the trace operator
$(\cdot)^{\mathsf{T}}$	denote the transpose operator
$(\cdot)^{\mathsf{H}}$	denote the Hermitian transpose operator
$\mathcal{CN}(\mathbf{X}, \mathbf{Y})$	denote the circular symmetric complex Gaussian distribution with mean matrix $\mathbf{X}$ and covariance matrix $\mathbf{Y}$
$\mathcal{U}(\cdot)$	distribution of a uniform random variable
$\mathcal{N}(x, \sigma)$	distribution of a normal random variable with mean x and variance σ^2
$\sim$	stands for "distributed as"
$\mathbb{E}[\cdot]$	denote expectation
$\mathbb{V}[\cdot]$	denote variance
$\|\mathbf{x}\|$	denote the Euclidean norm of complex vector $\mathbf{x}$
$\|z\|$	denote the magnitude of a complex number z
$\mathbb{I}[x]$	denote the indicator function; it returns 1 when x is true, and 0 otherwise

Appendix B

Math expressions and derivations

B.1 RACH Capacity: Approximations and Bounds

In this appendix, we derive some approximations and bounds for the system capacity, $c(R)$, defined in [2].

First, we recall that

$$1 - \frac{1}{x} < \log(x) < x - 1, \quad \text{for } x > 0. \tag{B.1}$$

From (B.1), it follows immediately that

$$R - 1 < \left[\log\left(\frac{R}{R-1} \right) \right]^{-1} < R. \tag{B.2}$$

Applying the inequalities in (B.2) to (2.3) we obtain

$$\ell_0(R) < c(R) < u(R), \tag{B.3}$$

where

$$\ell_0(R) \triangleq (R-1)\left(1 - \frac{1}{R}\right)^{R-1} = R\left(1 - \frac{1}{R}\right)^{R}, \tag{B.4}$$

$$u(R) \triangleq R\left(1 - \frac{1}{R}\right)^{R-2}. \tag{B.5}$$

From (B.4) and (B.5) for $R > 0$, it can be easily seen that

$$\ell_0(R) < \ell_1(R) < u(R), \tag{B.6}$$

where

$$\ell_1(R) \triangleq R\left(1 - \frac{1}{R}\right)^{R-1} \approx c(R). \tag{B.7}$$

Now, by observing that $(1 - 1/R)^R$ is increasing and tends to e^{-1}, and $(1 - 1/R)^{R-1}$ is decreasing and tends to e^{-1}, we can see that

$$\ell_0(R) < \ell_2(R) < \ell_1(R), \tag{B.8}$$

where

$$\ell_2(R) \triangleq \frac{R}{e}. \tag{B.9}$$

From the above observations, it can also be deduced that if R is sufficiently large, $\ell_0(R) \approx \ell_2(R) \approx \ell_1(R)$. Besides, by numerical evaluation we have verified that $\ell_1(R) < c(R)$.

Finally, combining the previous derivations we have

$$\ell_0(R) < \ell_2(R) < \ell_1(R) < c(R) < u(R) \tag{B.10}$$

and the approximations given in (2.4), i.e., $c(R) \approx \ell_1(R) \approx \ell_2(R)$. As can be seen in Table B.1, $\ell_1(R)$ provides an extremely accurate approximation, while $\ell_2(R)$, which is a simpler expression, can be considered as sufficiently accurate for all practical purposes (see also Fig. 2.6).

B.2 Joint PDF of the Number of Successful and Collided Preamble Transmissions

In this appendix, we present our devised expressions for computing the joint probability distribution function (PDF) of the number of successful and collided preamble transmissions within a random access opportunity (RAO). For this purpose, in Section B.2.1 we derive a closed-form expression and in

Table B.1: Accuracy of the Approximations and Bounds

		Rel. error (%)			
R	$c(R)$	$\ell_0(R)$	$\ell_2(R)$	$\ell_1(R)$	$u(R)$
10	3.8796	10.1248	5.1755	0.1386	10.9571
20	7.5496	5.0312	2.5427	0.0329	5.2285
30	11.2256	3.3472	1.6855	0.0144	3.4334
40	14.9030	2.5078	1.2605	0.0080	2.5559
50	18.5810	2.0050	1.0067	0.0051	2.0356
60	22.2593	1.6701	0.8380	0.0035	1.6913
70	25.9377	1.4311	0.7177	0.0026	1.4466

Section B.2.2 we devise a computationally efficient recursion useful when the number of contending UEs is large as it can occur when M2M communications are involved.

For the formulation, let us focus on a single RAO with r available preambles, and, to simplify the notation, let the random variable (r.v.) N be the number of contending UEs, i.e., the UEs that transmit a preamble randomly selected among the r available preambles. Denote by S the r.v. representing the number of preambles successfully transmitted, C the r.v. representing the number of collided preambles, U the r.v. representing the number of non-used preambles. If $N = n, (S, C) \in \mathcal{R}_n \triangleq \{(s, c) \in \mathbb{N}^2 \mid s + c \leq r, n - s \geq 2c\}$.

B.2.1 Closed-Form Expression

Before providing the closed-form expression we need to introduce the following lemma.

Lemma B.1. *Consider n UEs that are randomly assigned to c different preambles. Each UE is independently assigned to one preamble. Let* $\mathrm{F}(n, c)$ *denote the probability that at least two UEs are assigned to each preamble. Then,* $\mathrm{F}(n, c) = 0$ *if* $n < 2c$

and

$$\mathrm{F}(n,c) = 1 + \sum_{k=1}^{c-1} \left((-1)^k \binom{c}{k} \times \sum_{m=0}^{k} \binom{n}{m} \left(\frac{k}{c}\right)^m \left(1 - \frac{k}{c}\right)^{n-m} \frac{k!}{(k-m)!k^m} \right) \quad \text{(B.11)}$$

if $n \geq 2c$.

Proof. It is obvious that $\mathrm{F}(n,c) = 0$ if $n < 2c$, so in what follows we focus on the case $n \geq 2c$.

Consider a subset of $k < c$ preambles. The probability that exactly m UEs ($m = 1, \ldots, k$) are assigned to the preambles in this subset, with at most one UE per preamble, is given by

$$p_{m,k} = \binom{n}{m} \left(\frac{k}{c}\right)^m \left(1 - \frac{k}{c}\right)^{n-m} \frac{k}{k} \frac{k-1}{k} \cdots \frac{k-(m-1)}{k} = \binom{n}{m} \left(\frac{k}{c}\right)^m \left(1 - \frac{k}{c}\right)^{n-m} \frac{k!}{(k-m)!k^m}. \quad \text{(B.12)}$$

It is easy to check that the right-hand side of the expression above is also valid for $m = 0$.

Let A_k denote the event that less than two (i.e., one or none) UEs have been assigned to the k-th preamble. Then, we can write

$$\begin{aligned} \mathrm{F}(n,c) &= 1 - \Pr(A_1 \cup A_2 \cup \cdots \cup A_c) \\ &= 1 - \sum_{k=1}^{c} (-1)^{k-1} \sum_{1 \leq i_1 < \cdots < i_k \leq c} \Pr(\bigcap_{j=1}^{k} A_{i_j}) \\ &= 1 - \sum_{k=1}^{c-1} (-1)^{k-1} \sum_{1 \leq i_1 < \cdots < i_k \leq c} \sum_{m=0}^{k} p_{m,k} \\ &= 1 + \sum_{k=1}^{c-1} (-1)^k \binom{c}{k} \sum_{m=0}^{k} p_{m,k}. \end{aligned} \quad \text{(B.13)}$$

Above we have used the fact that $\Pr(\bigcap_{j=1}^{c} A_i) = 0$ since, as stated above, it is assumed that $n \geq 2c$. ■

Theorem B.2. *The conditional joint probability of having exactly $S = s$ successful preambles and $C = c$ collided preambles when $N = n$ UEs transmitted their preambles, $P_n(s,c) = \Pr(S = s, C = c, U = r - s - c \mid N = n)$, is given by*

$$P_n(s,c) = \begin{cases} \binom{r}{n}\frac{n!}{r^n}, & c = 0, s = n < r, \\ \binom{r}{c}\left(\frac{c}{r}\right)^n \mathrm{F}(n,c), & s = 0, 0 < c \leq c_{\max}, \\ \binom{r}{s,c,r-s-c}\binom{n}{s}\frac{s!}{r^s}\left(\frac{c}{r}\right)^{n-s} \times \mathrm{F}(n-s,c), & s > 0, c > 0, \\ 0, & \textit{otherwise}, \end{cases} \tag{B.14}$$

where $c_{\max} = \min\{r, \lfloor N/2 \rfloor\}$.

Proof. 1. $c = 0, s = n < r$

First, consider a fixed subset, $\mathcal{P}_s$, with $s = n$ preambles formed by the preambles that will be chosen by exactly one UE. Since each of the n UEs randomly chooses a preamble among the r available ones, the probability that each preamble in $\mathcal{P}_s$ is chosen by exactly one UE is given as

$$\frac{s}{r} \times \frac{s-1}{r} \times \cdots \times \frac{2}{r} \times \frac{1}{r} = \frac{s!}{r^n}. \tag{B.15}$$

Since the number of different subsets $\mathcal{P}_s$ is $\binom{r}{s}$, we have

$$P_n(s,c) = \binom{r}{s}\frac{s!}{r^n} = \binom{r}{n}\frac{n!}{r^n}. \tag{B.16}$$

2. $s = 0, c = 1, \ldots, c_{\max}$

Similarly to the previous case, consider a fixed subset, $\mathcal{P}_c$, with c preambles formed by the collided preambles (i.e., preambles that will be chosen by at least two UEs); there are $\binom{r}{c}$ different such subsets.

The probability that the n UEs choose a preamble in $\mathcal{P}_c$ (with no assumption on how the UEs distribute among the c preambles) is $(c/r)^n$.

Now, using the function F to account for the probability that the n UEs distribute among the c preambles in $\mathcal{P}_c$ in such a way that each

preamble is chosen by at least two UEs, we can write

$$P_n(s,c) = \binom{r}{c} \left(\frac{c}{r}\right)^n \mathrm{F}(n,c). \tag{B.17}$$

3. $s > 0, c > 0$

   Consider now two fixed (and disjoint) subsets, $\mathcal{P}_s$ and $\mathcal{P}_c$, with of s and c preambles, respectively, and corresponding to the successful and collided preambles; there are $\binom{r}{s,c,r-s-c}$ different ways to form these two subsets.

   Likewise, consider a fixed subset, $\mathcal{U}_s$, with s UEs corresponding to the UEs that will choose a preamble from $\mathcal{P}_s$; the remaining $n-s$ UEs will choose a preamble from $\mathcal{P}_c$. There are $\binom{n}{s}$ ways to choose the subset $\mathcal{U}_s$.

   Combining the reasoning used in the two previous cases, we can finally write

$$P_n(s,c) = \binom{r}{s,c,r-s-c} \binom{n}{s} \frac{s!}{r^s} \left(\frac{c}{r}\right)^{n-s} \mathrm{F}(n-s,c). \tag{B.18}$$

■

The conditional PDFs of S and C for a given N are the marginal probability distributions computed as

$$P_n(s) = \Pr(S = s \mid N = n) = \sum_{c=0}^{c_{\max}} P_n(s,c). \tag{B.19}$$

$$P_n(c) = \Pr(C = c \mid N = n) = \sum_{s=0}^{\min\{r,N\}} P_n(s,c). \tag{B.20}$$

We emphasize the simplicity of this formulation in comparison with that of the existing methods [22, 55, 99] used for the same purpose.

B.2.2 Recursion

In Theorem B.2, we presented a simple expression (B.14) to compute the conditional joint probability of having exactly $S = s$ successful preambles and $C = c$ collided preambles when $N = n$ UEs transmitted their preambles. We can also find in the literature another compact explicit expression to compute the probability of having $S = s$ successes out of $N = n$ transmissions on r preambles [100, 101]. However, the limitation of involving the computation of factorials of large numbers is present as in [22, 55, 99]. To overcome this limitation, we devise a recursive method for computing the conditional joint PDF of S and C for a given N as follows

$$\begin{aligned} P_n(s,c) =& \frac{r-(s-1+c)}{r} P_{n-1}(s-1,c) + \\ & \frac{s+1}{r} P_{n-1}(s+1,c-1) + \frac{c}{r} P_{n-1}(s,c), \end{aligned} \tag{B.21}$$

given the initial condition $P_0(0,0) = 1$ and $P_n(s,c) = 0$ if $(s,c) \notin \mathcal{R}_n$.

That is, from the distribution for the case with $n-1$ UEs, we obtain the distribution when a UE is added. The reasoning behind the recursive expression in (B.21) is as follows. Let $(s,c)_n$ represent the case in which n UEs have chosen their preambles so that s preambles were chosen by exactly one UE and c preambles were chosen by at least two UEs. Consider now the following three instances that can lead to s successfully transmitted preambles and c collided preambles when an extra UE chooses a preamble:

1. The first $n-1$ UEs have chosen their preambles in such a way that $s-1$ preambles were chosen by exactly one UE and c preambles were chosen by at least two UEs, i.e., we are in the case $(s-1,c)_{n-1}$. Then, the nth UE chooses one of the $r-(s-1+c)$ preambles that was not chosen by any of the previous $n-1$ UEs. Obviously, the probability of the nth UE making this selection is $(r-(s-1+c))/r$.
2. We start from $(s+1,c-1)_{n-1}$ and the nth UE chooses one of the $s+1$ preambles that was chosen by exactly one of the previous $n-1$ UEs; the probability of the latter is $(s+1)/r$.

3. We start from $(s,c)_{n-1}$ and the nth UE chooses one of the c preambles that was already chosen by more than one of the previous $n-1$ UEs; the probability of the latter is c/r.

B.3 Phase-Type Distribution

Consider a CTMC on a finite state-space $\mathcal{S} = \{0, 1, 2, \ldots, m\}$ where one state is absorbing and the remaining m states are transient. The random variable defined as the time to absorption is said to have a continuous PH distribution [102].

A PH distribution is uniquely given by the pair $(\boldsymbol{\alpha}; \boldsymbol{T})$, where $\boldsymbol{\alpha}$ is a m-dimensional row vector that defines the probabilities that the system starts at any of the transient states and meet $\sum_{i=0}^{m} \alpha_i = 1$; while T is a $m \times m$ matrix referred to as the *PH generator* that contains the transition rates between the transient states.

The infinitesimal generator for the CTMC can be written in block-matrix form as $\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{T} & \boldsymbol{t} \\ \boldsymbol{0} & 0 \end{bmatrix}$. Here, $\boldsymbol{0}$ is a $1 \times m$ row vector of zeros. The elements of the column vector $\boldsymbol{t} = [t_1, t_2, \ldots, t_m]'$ are the transition rates from the transient states to the absorbing state. The $m \times m$ sub-stochastic matrix $\boldsymbol{T}$ meets $\boldsymbol{t} = -\boldsymbol{T}\boldsymbol{e}$, where $\boldsymbol{e}$ is a column vector of ones of appropriate dimension.

It is known that $-(\boldsymbol{T}^{-1})_{ij}$ is the expected total time spent in phase j during the time until absorption, conditioned on the system starting at phase i [103, Theorem 2.4.3]. The elements of $-\boldsymbol{T}^{-1}$ are used to obtain the fractions of time the system spends at each of the m states until absorption. The interested reader is referred to [102–104] for further details and a comprehensive theoretical treatment of PH distributions.

Appendix C

Publications

C.1 Related with this dissertation

C.1.1 Journal

1. L. Tello-Oquendo, I. Leyva-Mayorga, V. Pla, J. Martinez-Bauset, J. R. Vidal, L. Guijarro, V. Casares-Giner, "Performance Analysis and Optimal Access Class Barring Parameter Configuration in LTE-A Networks with Massive M2M Traffic," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 4, pp. 3505–3520, Apr. 2018.

2. I. Leyva-Mayorga, L. Tello-Oquendo, V. Pla, J. Martinez-Bauset, V. Casares-Giner, "On the Accurate Performance Evaluation of the LTE-A Random Access Procedure and the Access Class Barring Scheme," *IEEE Transaction on Wireless Communications*, vol. 86, no. 12, pp. 7785–7799, Dec. 2017.

3. L. Tello-Oquendo, V. Pla, J. Martinez-Bauset, V. Casares-Giner, "Performance Analysis of Wireless Networks Based on Time-Scale Separation: A New Iterative Method," *Computer Communications*, vol. 86, pp. 40–48, Jul. 2016.

4. L. Tello-Oquendo, V. Pla, I. Leyva-Mayorga, J. Martinez-Bauset, V. Casares-Giner, L. Guijarro, "Efficient Random Access Channel Evaluation and Load Estimation in LTE-A with Massive MTC," submitted to *IEEE Transactions on Vehicular Technology*; Dec. 2017 (first submission), Jul. 2018 (second submission).

5. J. R. Vidal, L. Tello-Oquendo, V. Pla, L. Guijarro, "Collision-Avoidance Scheme to Enhance Extended Access Barring Performance in Massive Machine-Type Cellular Communications," to be submitted, 2018.

C.1.2 International conferences

1. L. Tello-Oquendo, D. Pacheco-Paramo, V. Pla, J. Martinez-Bauset, "Reinforcement Learning-Based ACB in LTE-A Networks for Handling Massive M2M and H2H Communications," IEEE International Conference on Communications (ICC), May 2018.

2. L. Tello-Oquendo, I. F. Akyildiz, S.-C. Lin, V. Pla, "SDN-Based Architecture for Providing Reliable Internet of Things Connectivity in 5G Systems," 17th Annual Mediterranean Ad Hoc Networking Workshop (Med-Hoc-Net), Jun. 2018.

3. L. Tello-Oquendo, J. R. Vidal, V. Pla, L. Guijarro, "Dynamic Access Class Barring Parameter Tuning in LTE-A Networks with Massive M2M Traffic," 17th Annual Mediterranean Ad Hoc Networking Workshop (Med-Hoc-Net), Jun. 2018.

4. L. Tello-Oquendo, I. F. Akyildiz, S.-C. Lin, V. Pla, "A Software-Defined Networking based Architecture for QoS-Aware IoT Communication in 5G Systems," 2nd International Balkan Conference on Communications and Networking (BalkanCom), Jun. 2018.

5. L. Tello-Oquendo, J. R. Vidal, V. Pla, J. Martinez-Bauset, "Extended Access Barring in Cellular-based Machine Type Communications: Practical Implementation and Impact of Paging Timing," Workshop on Inno-

vation on Information and Communication Technologies (WIICT), Jul. 2018.

6. I. Leyva-Mayorga, L. Tello-Oquendo, V. Pla, J. Martinez-Bauset, V. Casares-Giner, "On the Accurate Performance Evaluation of the LTE-A Random Access Procedure," IEEE Global Communications Conference (GLOBECOM), Dec. 2017.

7. L. Tello-Oquendo, S.-C. Lin, I. F. Akyildiz, V. Pla, J. Martinez-Bauset, V. Casares-Giner, "A Software-Defined Networking Architecture for 5G Internet-of-Things Communication," Workshop on Innovation on Information and Communication Technologies (WIICT), pp. 1–11, 2017.

8. I. Leyva Mayorga, L. Tello-Oquendo, V. Pla, J. Martinez-Bauset, "Collision handling in the LTE-A random access procedure: common assumptions and their impact on performance," Workshop on Innovation on Information and Communication Technologies (WIICT), pp. 19–30, 2017.

9. L. Tello-Oquendo, I. F. Akyildiz, S.-C. Lin, V. Pla, J. Martinez-Bauset, V. Casares-Giner, "A SDN-Based Architecture for Providing Internet of Things Connectivity in 5G Systems," 3rd Open International Workshop on Elastic Networks: 5G, Oct. 2017.

10. I. Leyva-Mayorga, L. Tello-Oquendo, V. Pla, J. Martinez-Bauset, V. Casares-Giner,"Performance Analysis of Access Class Barring for Handling Massive M2M Traffic in LTE-A Networks," IEEE International Conference on Communications (ICC), pp. 1–6, May 2016.

11. L. Tello-Oquendo, I. Leyva-Mayorga, V. Pla, J. Martinez-Bauset, V. Casares-Giner, "Estimating the Number of Contending Users for a Single Random Access in LTE-A networks: The Baseline for Designing Congestion Control Schemes at the Evolved Node B," Workshop on Innovation on Information and Communication Technologies (WIICT), pp. 32–41, 2016.

12. L. Tello-Oquendo, I. Leyva-Mayorga, V. Pla, J. Martinez-Bauset, V. Casares-Giner, “Analysis of LTE-A Random Access Procedure: A Foundation to Propose Mechanisms for Managing the M2M Massive Access in Wireless Cellular Networks,” Workshop on Innovation on Information and Communication Technologies (WIICT), pp. 95–104, 2015.

13. L. Tello-Oquendo, V. Pla, J. Martinez-Bauset, "Performance Analysis of Telecommunication Systems based on Time-Scale Separation," Workshop on Innovation on Information and Communication Technologies (WIICT), pp. 189–198, 2014.

14. L. Tello-Oquendo, V. Pla, J. Martinez-Bauset, "Approximate analysis of wireless systems based on time-scale decomposition," IFIP Wireless Days (WD), pp. 1–6, Nov. 2013.

C.2 Other publications

C.2.1 Journal

1. I. Leyva-Mayorga, M. A. Rodriguez-Hernandez, V. Pla, J. Martinez-Bauset, L. Tello-Oquendo, “Adaptive access class barring for efficient mMTC,” submitted, Jul. 2018.

C.2.2 International conferences

1. D. Ghose, L. Tello-Oquendo, F. Y. Li, V. Pla, “Lightweight Relay Selection in Multi-hop Wake-up Radio Enabled IoT Networks,” accepted, IEEE Global Communications Conference (GLOBECOM), 2018.

2. L. Tello-Oquendo, V. Casares-Giner, V. Pla, J. Martinez-Bauset, “On the Splitting Algorithms for Efficient Resource Allocation in LTE-A Networks: Some Insights,” 4th Open International Workshop on Elastic Networks, Jan. 2018.

3. L. Tello-Oquendo, V. Pla, J. Martinez-Bauset, V. Casares-Giner, "Efficient Access Control for Managing Massive M2M Communications over Cellular Networks," IEEE ComSoc Summer Conference, Jul. 2017.

4. I. Leyva-Mayorga, L. Tello-Oquendo, V. Pla, J. Martinez-Bauset, V. Casares-Giner, "On the impact of the capture effect in the LTE-A random access channel," Workshop on Innovation on Information and Communication Technologies (WIICT), 2017.

5. I. Leyva-Mayorga, M. A. Rodriguez-Hernandez, V. Pla, J. Martinez-Bauset, L. Tello-Oquendo, "An Adaptive Access Class Barring Scheme for Handling Massive M2M Communications in LTE-A," 23rd European Wireless Conference (EW), pp. 143–148, 2017.

6. I. Leyva-Mayorga, M. A. Rodriguez-Hernandez, V. Pla, J. Martinez-Bauset, L. Tello-Oquendo, V. Casares-Giner, "An Adaptive Access Class Barring Scheme for Handling Massive M2M Communications in LTE-A," International Workshop on Elastic Networks Design and Optimization (ELASTICNETS), 2017.

7. I. Leyva-Mayorga, L. Tello-Oquendo, V. Pla, J. Martinez-Bauset, V. Casares-Giner, "A dynamic access class barring method to avoid congestion in LTE-A networks with massive M2M traffic," Workshop on Innovation on Information and Communication Technologies (WIICT), pp. 60–69, 2016.

C.2.3 National conferences

1. V. Casares-Giner, L. Tello-Oquendo, V. Pla, J. Martinez-Bauset, "On the Dimensionality Reduction of Markov Chains for Networks Modeling," XIII Jornadas de Ingeniería Telemática, pp. 107–109, Sep. 2017.

2. L. Tello-Oquendo, I. Leyva-Mayorga, V. Pla, J. Martinez-Bauset, "Estimating the Number of Contending Users at the Random Access Channel in LTE-A networks to Design Congestion Control Schemes for Pro-

viding Machine-Type Communications," III Meeting of Ph.D. students, Universitat Politècnica de València, 2016.

Appendix D

Research projects

This work has been developed in the framework of the following research projects:

- *Cooperation and Opportunism in Heterogeneous Wireless Access Networks*, COHWAN, TIN2010-21378-C02-02. January 1, 2011 – June 30, 2014.
- *Platform of Services for Smart Cities with Dense Machine-to-Machine Networks*, PLASMA, TIN2013-47272-C2-1-R. January 1, 2014 – January 1, 2018.
- *New Paradigms of Elastic Networks for a World Radically Based on Cloud and Fog Computing*, Elastic Networks, TEC2015-71932-REDT. December 1, 2015 – November 30, 2018.

Likewise, the author thanks the support from the Ecuadorian Government through the *Secretaría Nacional de Educación Superior, Ciencia, Tecnología e Innovación (SENESCYT)* and its *Programa de Estudios de Cuarto Nivel de Formación Académica en el Exterior "Convocatoria Abierta 2012" Segunda Fase*.

Finally, we thank the support from *Programa de Ayudas de Investigación y Desarrollo (PAID)*, Universitat Politècnica de València.

Bibliography

[1] 3GPP, *TS 36.211, Physical Channels and Modulation*, Sep 2017.

[2] 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, Jun 2014.

[3] I. F. Akyildiz, P. Wang, and S.-C. Lin, "SoftAir: A software defined networking architecture for 5G wireless systems," *Computer Networks*, vol. 85, pp. 1 – 18, 2015.

[4] 3GPP, *TS 22.011, V15.1.0, Service Accessibility*, Sep 2017.

[5] ——, *TR 37.868, Study on RAN Improvements for Machine Type Communications*, Sep 2011.

[6] Ericsson. (2017, Nov.) Ericsson mobility report. [Online]. Available: https://www.ericsson.com/mobility-report

[7] Cisco. (2017, Mar.) Cisco visual networking index (VNI): Global mobile data traffic forecast update, 2016-2021. [Online]. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html

[8] Z. Dawy, W. Saad, A. Ghosh, J. G. Andrews, and E. Yaacoub, "Toward massive machine type cellular communications," *IEEE Wireless Communications*, vol. 24, no. 1, pp. 120–128, February 2017.

[9] I. F. Akyildiz, S. Nie, S.-C. Lin, and M. Chandrasekaran, "5G roadmap: 10 key enabling technologies," *Computer Networks*, vol. 106, pp. 17 – 48, 2016.

[10] 3GPP, *TS 23.682, Architecture enhancements to facilitate communications with packet data networks and applications*, Mar 2016.

[11] ——, *TS 22.368, Service Requirements for Machine-Type Communications*, Mar 2017.

[12] I. S. Association. Internet of things. [Online]. Available: http://standards.ieee.org/innovate/iot/

[13] P. K. Verma, R. Verma, A. Prakash, A. Agrawal, K. Naik, R. Tripathi, M. Alsabaan, T. Khalifa, T. Abdelkader, and A. Abogharaf, "Machine-to-Machine (M2M) communications: A survey ," *J. Netw. Comput. Appl.*, vol. 66, pp. 83 – 105, 2016.

[14] Y. Mehmood, C. Görg, M. Muehleisen, and A. Timm-Giel, "Mobile M2M communication architectures, upcoming challenges, applications, and future directions," *EURASIP J. Wirel. Commun. Netw.*, vol. 2015, no. 1, pp. 1–37, 2015.

[15] 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, May 2015.

[16] A. Lo, Y. Law, and M. Jacobsson, "A cellular-centric service architecture for machine-to-machine (M2M) communications," *IEEE Wireless Commun. Mag.*, vol. 20, no. 5, pp. 143–151, 2013.

[17] S.-Y. Lien, K.-C. Chen, and Y. Lin, "Toward ubiquitous massive accesses in 3GPP machine-to-machine communications," *IEEE Commun. Mag.*, vol. 49, no. 4, pp. 66–74, 2011.

[18] 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.

[19] L. Ferdouse, A. Anpalagan, and S. Misra, "Congestion and overload control techniques in massive M2M systems: a survey," *Trans. Emerg. Telecommun. Technol.*, vol. 25, no. 3, pp. 1–17, Mar 2015.

[20] 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, Jan 2014.

[21] 3GPP, *TS 36.331, Radio Resource Control (RRC), Protocol specification*, Sep 2017.

[22] 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.

[23] 3GPP, *TS 36.321, Medium Access Control (MAC) Protocol Specification*, Sep 2017.

[24] 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 International Conference on Communications (ICC)*, May 2016, pp. 1–6.

[25] C. Y. Oh, D. Hwang, and T. J. Lee, "Joint Access Control and Resource Allocation for Concurrent and Massive Access of M2M Devices," *IEEE Trans. Wireless Commun.*, vol. 14, no. 8, pp. 4182–4192, Aug 2015.

[26] M. Tavana, V. Shah-Mansouri, and V. W. S. Wong, "Congestion control for bursty M2M traffic in LTE networks," in *Proc. IEEE International Conference on Communications (ICC)*, Jun 2015, pp. 5815–5820.

[27] O. Arouk and A. Ksentini, "General Model for RACH Procedure Performance Analysis," *IEEE Commun. Lett.*, vol. 20, no. 2, pp. 372–375, Feb 2016.

[28] 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, Nov 2014.

[29] G.-Y. Lin, S.-R. Chang, and H.-Y. Wei, "Estimation and Adaptation for Bursty LTE Random Access," *IEEE Trans. Veh. Technol.*, vol. 65, no. 4, pp. 2560–2577, 2016.

[30] J. J. Nielsen, D. M. Kim, G. C. Madueno, N. K. Pratas, and P. Popovski, "A tractable model of the LTE access reservation procedure for machine-type communications," in *Proc. IEEE Global Communications Conference (GLOBECOM)*, Dec 2015, pp. 1–6.

[31] 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.

[32] A. Biral, M. Centenaro, A. Zanella, L. Vangelista, and M. Zorzi, "The challenges of M2M massive access in wireless cellular networks," *Digit. Commun. Netw.*, vol. 1, no. 1, pp. 1–19, 2015.

[33] 3GPP, *TS 36.213, Physical layer procedures*, Dec 2014.

[34] ——, *TR 36.912, Feasibility study for Further Advancements for E-UTRA*, Mar 2017.

[35] J. E. Wieselthier, A. Ephremides, and L. A. Michaels, "An exact analysis and performance evaluation of framed ALOHA with capture," *IEEE Trans. Commun.*, vol. 37, no. 2, pp. 125–137, Feb 1989.

[36] D. C. Chu, "Polyphase codes with good periodic correlation properties," *IEEE Trans. Inf. Theory*, vol. 18, 1972.

[37] M. M. Mansour, "Optimized architecture for computing Zadoff-Chu sequences with application to LTE," in *Proc. IEEE Global Telecommunications Conference (GLOBECOM)*, vol. 2, no. 1, 2009.

[38] F. A. P. De Figueiredo, F. S. Mathilde, F. A. C. M. Cardoso, R. M. Vilela, and J. P. Miranda, "Efficient frequency domain zadoff-chu generator with application to LTE and LTE-A systems," in *Proc. International Telecommunications Symposium, (ITS)*, 2014, pp. 1–5.

[39] C. L. Taylor, D. Nolan, and S. Wainberg, "Priority capabilities in LTE supporting national security and emergency preparedness next generation network priority services," in *Proc. IEEE International Conference on Technologies for Homeland Security (HST)*, Nov 2013, pp. 584–588.

[40] D. Nolan, S. Wainberg, J. R. Wullert, and A. R. Ephrath, "National security and emergency preparedness communications: Next generation priority services," in *Proc. IEEE International Conference on Technologies for Homeland Security (HST)*, Nov 2013, pp. 106–112.

[41] L. Segura, "Access control for M2M devices," Aug. 18 2011, US Patent App. 13/028,093.

[42] H. Thomsen, N. K. Pratas, v. Stefanovic, and P. Popovski, "Code-expanded radio access protocol for machine-to-machine communications," *Trans. Emerg. Telecommun. Technol.*, vol. 24, no. 4, pp. 355–365, 2013.

[43] M. Condoluci, M. Dohler, G. Araniti, A. Molinaro, and J. Sachs, "Enhanced Radio Access and Data Transmission Procedures Facilitating Industry-Compliant Machine-Type Communications over LTE-Based 5G Networks," *IEEE Wireless Commun. Mag.*, vol. 23, no. 1, pp. 56–63, 2016.

[44] M. Condoluci, G. Araniti, M. Dohler, A. Iera, and A. Molinaro, "Virtual code resource allocation for energy-aware MTC access over 5G systems," *Ad Hoc Netw.*, vol. 43, pp. 3–15, 2016.

[45] M. Shirvanimoghaddam, M. Dohler, and S. J. Johnson, "Massive Multiple Access Based on Superposition Raptor Codes for Cellular M2M Communications," *IEEE Trans. Wireless Commun.*, vol. 16, no. 1, pp. 307–319, Jan 2017.

[46] M. S. Shafiq, L. Ji, A. X. Liu, J. Pang, A. Venkataraman, and J. Wang, "A First Look at Cellular Network Performance during Crowded Events," in *ACM SIGMETRICS/international conference on Measurement and modeling of computer systems*, June 2013.

[47] M. S. Shafiq, J. Erman, L. Ji, A. Liu, J. Pang, and J. Wang, "Understanding the Impact of Network Dynamics on Mobile Video User Engagement," in *ACM SIGMETRICS/international conference on Measurement and modeling of computer systems*, June 2014.

[48] H. Zang and J. Bolot, "Mining call and mobility data to improve paging efficiency in cellular networks," in *Proceedings of the 13th annual ACM international conference on Mobile computing and networking*, Sep 2007.

[49] 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, April 2018.

[50] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press Cambridge, 1998, vol. 1, no. 1.

[51] H. Kim, S. s. Lee, and S. Lee, "Dynamic extended access barring for improved M2M communication in LTE-A networks," in *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Oct 2017, pp. 2742–2747.

[52] R.-H. Hwang, C.-F. Huang, H.-W. Lin, and J.-J. Wu, "Uplink access control for machine-type communications in lte-a networks," *Personal and Ubiquitous Computing*, vol. 20, no. 6, pp. 851–862, Nov 2016.

[53] C. M. Chou, C. Y. Huang, and C.-Y. Chiu, "Loading prediction and barring controls for machine type communication," in *2013 IEEE International Conference on Communications (ICC)*. IEEE, jun 2013, pp. 5168–5172.

[54] A. Lo, Y.-W. Law, M. Jacobsson, and M. Kucharzak, "Enhanced LTE-advanced random-access mechanism for massive machine-to-machine (M2M) communications," 2011.

[55] 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, Dec 2016.

[56] L. M. Bello, P. Mitchell, and D. Grace, "Application of Q-Learning for RACH Access to Support M2M Traffic over a Cellular Network," in *Proc. 20th European Wireless Conference*, May 2014.

[57] J. Moon and Y. Lim, "A Reinforcement Learning Approach to Access Management in Wireless Cellular Networks," *Wireless Communications and Mobile Computing*, May 2017.

[58] C. J. Watkins and P. Dayan, "Technical note: Q-learning," *Machine Learning*, vol. 8, no. 3, pp. 279–292, May 1992.

[59] J. N. Tsitsiklis, "Asynchronous stochastic approximation and Q-learning," *Machine Learning*, vol. 16, no. 3, pp. 185–202, Sep 1994.

[60] Telecomitalia. (2016, Nov.) Telecom italia: Big data challenge. [Online]. Available: http://www.telecomitalia.com/tit/en/innovazione/archivio/big-data-challenge-2015.html

[61] Nokia, "Mobile Broadband solutions for Mass Events," Nokia, Tech. Rep., 2014.

[62] U. Raza, P. Kulkarni, and M. Sooriyabandara, "Low power wide area networks: An overview," *IEEE Commun. Surveys Tuts.*, vol. PP, no. 99, 2017.

[63] R. S. Sinha, Y. Wei, and S.-H. Hwang, "A survey on LPWA technology: LoRa and NB-IoT," *ICT Express*, 2017.

[64] I. F. Akyildiz, S.-C. Lin, and P. Wang, "Wireless software-defined networks (W-SDNs) and network function virtualization (NFV) for 5G cellular systems: An overview and qualitative evaluation," *Computer Networks*, vol. 93, Part 1, pp. 66 – 79, 2015.

[65] I. S. Association. IEEE Standard for Low-Rate Wireless Networks. [Online]. Available: http://www.ieee802.org/15/pub/TG4.html

[66] M. C. Vuran and I. F. Akyildiz, "Error control in wireless sensor networks: A cross layer analysis," *IEEE/ACM Transactions on Networking*, vol. 17, no. 4, pp. 1186–1199, Aug 2009.

[67] C. Han, J. M. Jornet, E. Fadel, and I. F. Akyildiz, "A cross-layer communication module for the internet of things," *Computer Networks*, vol. 57, no. 3, pp. 622 – 633, 2013.

[68] W. Ye, J. Heidemann, and D. Estrin, "Medium access control with coordinated adaptive sleeping for wireless sensor networks," *IEEE/ACM Transactions on Networking*, vol. 12, no. 3, pp. 493–506, June 2004.

[69] E. Lähetkangas, K. Pajukoski, J. Vihriälä, G. Berardinelli, M. Lauridsen, E. Tiirola, and P. Mogensen, "Achieving low latency and energy consumption by 5G TDD mode optimization," in *2014 IEEE International Conference on Communications Workshops (ICC)*, June 2014, pp. 1–6.

[70] C. She, C. Yang, and T. Q. S. Quek, "Cross-layer transmission design for tactile internet," in *2016 IEEE Global Communications Conference (GLOBECOM)*, Dec 2016, pp. 1–6.

[71] M. R. Akdeniz, Y. Liu, M. K. Samimi, S. Sun, S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter wave channel modeling and cellular capacity evaluation," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1164–1179, June 2014.

[72] M. D. Renzo, "Stochastic geometry modeling and analysis of multitier millimeter wave cellular networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 9, pp. 5038–5057, Sept 2015.

[73] S.-C. Lin and I. F. Akyildiz, "Dynamic base station formation for solving NLOS problem in 5G millimeter-wave communication," in *IEEE Conference on Computer Commun. (INFOCOM)*, may 2017, pp. 1–9.

[74] J. Jose, A. Ashikhmin, T. L. Marzetta, and S. Vishwanath, "Pilot contamination and precoding in multi-cell TDD systems," *IEEE Trans. Wireless Commun.*, vol. 10, no. 8, pp. 2640–2651, August 2011.

[75] V. N. Ha, L. B. Le, and N. D. Dao, "Coordinated multipoint transmission design for cloud-rans with limited fronthaul capacity constraints," *IEEE Trans. Veh. Technol.*, vol. 65, no. 9, pp. 7432–7447, Sept 2016.

[76] N. Tadayon and S. Aissa, "Modeling and analysis framework for multi-interface multi-channel cognitive radio networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 2, pp. 935–947, Feb 2015.

[77] W. Zhang, M. Suresh, R. Stoleru, and H. Chenji, "On modeling the coexistence of 802.11 and 802.15.4 networks for performance tuning," *IEEE Trans. Wireless Commun.*, vol. 13, no. 10, pp. 5855–5866, Oct 2014.

[78] G. G. Yin and Q. Zhang, *Discrete-time Markov chains: two-time-scale methods and applications*. Springer, 2006, vol. 55.

[79] W. J. Stewart, *Probability, Markov chains, queues, and simulation: the mathematical basis of performance modeling*. Princeton University Press, 2009.

[80] C. Politis, "Managing the radio spectrum," *Vehicular Technology Magazine*, vol. 4, no. 1, pp. 20–26, 2009.

[81] Y.-C. Liang, K.-C. Chen, G. Li, and P. Mahonen, "Cognitive radio networking and communications: an overview," *IEEE Trans. Veh. Technol.*, vol. 60, no. 7, pp. 3386–3407, Sept 2011.

[82] J. W. Roberts, "Internet traffic, QoS, and pricing," *Proceedings of the IEEE*, vol. 92, no. 9, pp. 1389–1399, 2004.

[83] W. Song, H. Jiang, W. Zhuang, and X. Shen, "Resource management for QoS support in cellular/WLAN interworking," *Network, IEEE*, vol. 19, no. 5, pp. 12–18, 2005.

[84] E. Wong and C. Foh, "Analysis of cognitive radio spectrum access with finite user population," *IEEE Communications Letters*, vol. 13, no. 5, pp. 294–296, 2009.

[85] J. Martinez-Bauset, V. Pla, J. Vidal, and L. Guijarro, "Approximate analysis of cognitive radio systems using time-scale separation and its accuracy," *IEEE Communications Letters*, vol. 17, no. 1, pp. 35–38, Jan. 2013.

[86] Y. Huang, K. Ko, and M. Zukerman, "A generalized quasi-stationary approximation for analysis of an integrated service system," *IEEE Communications Letters*, vol. 16, no. 11, pp. 1884–1887, Nov. 2012.

[87] L. Jiao, E. Song, V. Pla, and F. Li, "Capacity upper bound of channel assembling in cognitive radio networks with quasistationary primary user activities," *IEEE Trans. Veh. Technol.*, vol. 62, no. 4, pp. 1849–1855, May 2013.

[88] S. Liu and J. Virtamo, "Performance analysis of wireless data systems with a finite population of mobile users," in *Proceedings of the 19th International Teletraffic Congress ITC 19*, 2005, pp. 1295–1304.

[89] O. J. Boxma, A. F. Gabor, R. Núñez-Queija, and H.-P. Tan, "Performance analysis of admission control for integrated services with minimum rate guarantees," in *Proceedings of NGI'06*, 2006, pp. 41–47.

[90] L. Jiao, F. Y. Li, and V. Pla, "Modeling and performance analysis of channel assembling in multichannel cognitive radio networks with spectrum adaptation," *IEEE Trans. Veh. Technol.*, vol. 61, no. 6, pp. 2686–2697, 2012.

[91] L. Jiao, I. Balapuwaduge, F. Li, and V. Pla, "On the performance of channel assembling and fragmentation in cognitive radio networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 10, pp. 5661–5675, Oct 2014.

[92] L. Tello-Oquendo, V. Pla, and J. Martinez-Bauset, "Approximate analysis of wireless systems based on time-scale decomposition," in *Wireless Days (WD), 2013 IFIP*, Nov 2013, pp. 1–6.

[93] H. Al-Mahdi, M. A. Kalil, F. Liers, and A. Mitschele-Thiel, "Increasing spectrum capacity for ad hoc networks using cognitive radios: an analytical model," *IEEE Communications Letters*, vol. 13, no. 9, pp. 676–678, Oct. 2009.

[94] J. Peha, "Sharing spectrum through spectrum policy reform and cognitive radio," *Proceedings of the IEEE*, vol. 97, no. 4, pp. 708–719, 2009.

[95] P. Courtois, "Decomposability, instabilities, and saturation in multiprogramming systems," *Communications of the ACM*, vol. 18, no. 7, pp. 371–377, 1975.

[96] F. Hubner and P. Tran-Gia, "Quasi-stationary analysis of a finite capacity asynchronous multiplexer with modulated deterministic input," *ITC-13, Copenhagen*, 1991.

[97] V. Alexiades and A. D. Solomon, *Mathematical modeling of melting and freezing processes*. Taylor & Francis, 1993.

[98] D. P. Heyman and M. J. Goldsmith, "Comparisons between aggregation/disaggregation and a direct algorithm for computing the stationary probabilities of a markov chain," *ORSA Journal on Computing*, vol. 7, no. 1, pp. 101–108, 1995.

[99] O. Arouk, A. Ksentini, and T. Taleb, "How accurate is the RACH procedure model in LTE and LTE-A?" in *2016 International Wireless Communications and Mobile Computing Conference (IWCMC)*, Sept 2016, pp. 61–66.

[100] I. Pountourakis and E. Sykas, "Analysis, stability and optimization of Aloha-type protocols for multichannel networks," *Computer Communications*, vol. 15, no. 10, pp. 619 – 629, 1992.

[101] W. Yue and Y. Matsumoto, *Performance analysis of multi-channel and multi-traffic on wireless communication networks*. Springer Science & Business Media, 2007.

[102] M. Neuts, *Matrix-geometric Solutions in Stochastic Models: An Algorithmic Approach*. The Johns Hopkins University Press, 1981.

[103] G. Latouche and V. Ramaswami, *Introduction to Matrix Analytic Methods in Stochastic Modeling*. ASA-SIAM, 1999.

[104] A. S. Alfa, *Queueing Theory for Telecommunications: Discrete Time Modelling of a Single Node System*. Springer, 2010.