Congestion Control for Massive Machine-Type Communications: Distributed and Learning-Based Approaches

Manal El Tanab

A Thesis
In the Department
of
Electrical and Computer Engineering

Presented in Partial Fulfillment of the Requirements
For the Degree of
Doctor of Philosophy (Electrical and Computer Engineering) at
Concordia University
Montréal, Québec, Canada

June, 2021

CONCORDIA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

This is to certify that the thesis prepared

By: Manal El Tanab

Entitled: **Congestion Control for Massive Machine-Type Communications: Distributed and Learning-Based Approaches**

and submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy (Electrical and Computer Engineering)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

Chair

Dr. Kudret Demirli

External Examiner

Dr. Cheng Li

External to Program

Dr. Amr Youssef

Examiner

Dr. Dongyu Qiu

Examiner

Dr. Reza Soleymani

Thesis Supervisor

Dr. Walaa Hamouda

Approved by:

Dr. Wei-Ping Zhu, Graduate Program Director

July 5, 2021

Dr. Mourad Debbabi, Dean
Gina Cody School of Engineering and Computer Science

Abstract

Congestion Control for Massive Machine-Type Communications: Distributed and Learning-Based Approaches

Manal El Tanab, PhD

Concordia University, 2021

The Internet of things (IoT) is going to shape the future of wireless communications by allowing seamless connections among wide range of everyday objects. Machine-to-machine (M2M) communication is known to be the enabling technology for the development of IoT. With M2M, the devices are allowed to interact and exchange data without or with little human intervention. Recently, M2M communication, also referred to as machine-type communication (MTC), has received increased attention due to its potential to support diverse applications including eHealth, industrial automation, intelligent transportation systems, and smart grids.

M2M communication is known to have specific features and requirements that differ from that of the traditional human-to-human (H2H) communication. As specified by the Third Generation Partnership Project (3GPP), MTC devices are inexpensive, low power, and mostly low mobility devices. Furthermore, MTC devices are usually characterized by infrequent, small amount of data, and mainly uplink traffic. Most importantly, the number of MTC devices is expected to highly surpass that of H2H devices. Smart cities are an example of such a mass-scale deployment. These features impose various challenges related to efficient energy management, enhanced coverage and diverse quality of service (QoS) provisioning, among others.

The diverse applications of M2M are going to lead to exponential growth in M2M traffic. Associating with M2M deployment, a massive number of devices are expected to access the wireless network concurrently. Hence, a network congestion is likely to occur. Cellular networks have been recognized as excellent candidates for M2M support. Indeed, cellular networks are mature, well-established networks with ubiquitous coverage and reliability which allows cost-effective deployment of M2M communications. However, cellular networks were originally designed for human-centric services with high-cost devices and ever-increasing rate requirements. Additionally, the conventional random access (RA) mechanism used in Long Term Evolution-Advanced (LTE-A) networks lacks the capability of handling such an enormous number of access attempts expected from massive MTC. Particularly, this RA technique

acts as a performance bottleneck due to the frequent collisions that lead to excessive delay and resource wastage. Also, the lengthy handshaking process of the conventional RA technique results in highly expensive signaling, specifically for M2M devices with small payloads. Therefore, designing an efficient medium access schemes is critical for the survival of M2M networks.

In this thesis, we study the uplink access of M2M devices with a focus on overload control and congestion handling. In this regard, we mainly provide two different access techniques keeping in mind the distinct features and requirements of MTC including massive connectivity, latency reduction, and energy management. In fact, full information gathering is known to be impractical for such massive networks of tremendous number of devices. Hence, we assure to preserve the low complexity, and limited information exchange among different network entities by introducing distributed techniques. Furthermore, machine learning is also employed to enhance the performance with no or limited information exchange at the decision maker. The proposed techniques are assessed via extensive simulations as well as rigorous analytical frameworks.

First, we propose an efficient distributed overload control algorithm for M2M with massive access, referred to as M2M-OSA. The proposed algorithm can efficiently allocate the available network resources to massive number of devices within relatively small, and bounded contention time and with reduced overhead. By resolving collisions, the proposed algorithm is capable of achieving full resources utilization along with reduced average access delay and energy saving. For Beta-distributed traffic, we provide analytical evaluation for the performance of the proposed algorithm in terms of the access delay, total service time, energy consumption, and blocking probability. This performance assessment accounted for various scenarios including slightly, and seriously congested cases, in addition to finite and infinite retransmission limits for the devices. Moreover, we provide a discussion of the non-ideal situations that could be encountered in real-life deployment of the proposed algorithm supported by possible solutions. For further energy saving, we introduced a modified version of M2M-OSA with traffic regulation mechanism.

In the second part of the thesis, we adopt a promising alternative for the conventional random access mechanism, namely fast uplink grant. Fast uplink grant was first proposed by the 3GPP for latency reduction where it allows the base station (BS) to directly schedule the MTC devices (MTDs) without receiving any scheduling requests. In our work, to handle the major

challenges associated to fast uplink grant namely, active set prediction and optimal scheduling, both non-orthogonal multiple access (NOMA) and learning techniques are utilized. Particularly, we propose a two-stage NOMA-based fast uplink grant scheme that first employs multi-armed bandit (MAB) learning to schedule the fast grant devices with no prior information about their QoS requirements or channel conditions at the BS. Afterwards, NOMA facilitates the grant sharing where pairing is done in a distributed manner to reduce signaling overhead. In the proposed scheme, NOMA plays a major role in decoupling the two major challenges of fast grant schemes by permitting pairing with only active MTDs. Consequently, the wastage of the resources due to traffic prediction errors can be significantly reduced. We devise an abstraction model for the source traffic predictor needed for fast grant such that the prediction error can be evaluated. Accordingly, the performance of the proposed scheme is analyzed in terms of average resource wastage, and outage probability. The simulation results show the effectiveness of the proposed method in saving the scarce resources while verifying the analysis accuracy. In addition, the ability of the proposed scheme to pick quality MTDs with strict latency is depicted.

Acknowledgments

First and foremost, it is by the grace of Allah (God), the Most Gracious and the Most Merciful, that this thesis has been completed. Indeed, all praise and thanks are due to Allah for His countless blessings and for giving me the strength and the perseverance to accomplish my dream.

After quite a long journey, here comes the time to finally finish my thesis and express my sincere gratitude and appreciation to my supervisor Prof. Walaa Hamouda. Professor Hamouda is such a great supporter that guides me through both ups and downs. Even during the most challenging moments throughout my PhD, he believed in me and gave me insightful advices that were crucial in my academic success. Professor Hamouda is a caring person that cares about you not only technically but also personally and socially. Literally, this kind of advisors is the one that pushes you towards success. Thank you professor for your vision, continuous guidance, and patience that allow this research to be completed. Without your valuable support, encouragement, and flexibility, this thesis would not have been possible.

I would like to express my genuine appreciation to my committee members: Dr. Cheng Li, Dr. Amr Youssef, Dr. Dongyu Qiu and Dr. M Reza Soleymani for the time and effort spent in reading my thesis, taking part in my PhD dissertation defense committee, and providing me with valuable comments and suggestions.

I am sincerely thankful to Concordia Graduate Student Support Program (GSSP) for partially funding my PhD's studies.

My ultimate thanks is dedicated to my beloved mother, Shereen Eladl, for the extraordinary efforts she has exerted to help me reach this important life milestone. I would also like to extend my sincere thanks to my sister, Marwa Ezzat, for her unconditional love and continuous support throughout my life. Without their love, care, support and prayers, I would never have become who I am today. I am wholeheartedly thankful to my late father, Ezzat Eltanab, who was my first role model and inspiration. I am also extremely grateful to Dr. Samia Elhakim, my mother-in-law, and Dr. Khairy Helwa, my father-in-law, for their endless support, love, and sincere prayers. Last but not least, special thanks goes to my beloved family. My dearest husband, Mohamed Helwa, I cannot thank you enough for all your support. Your suggestions helped me a lot, you inspired me and gave me new directions. Thank you for standing by my side when time gets hard. Thank you for always being there when there is no one. For my adorable daughters, Maryam and Maleeka, thank you for filling my life with love and hope. To

my husband and daughters, I deeply appreciate your sacrifices and patience that made my PhD journey possible.

To my dearest husband, Mohamed Helwa,

To my adorable daughters, Maryam and Maleeka,

To my beloved mother, Shereen Eladl, and my sister, Marwa,

To the souls of my father, Ezzat Eltanab, and my grandmother Zeinab Sabra.

Contents

List of Figures

List of Tables

List of Acronyms

ACB	Access Class Barring
AP	Access Probability
BIP	Binary Integer Programming
BS	Base Station
CAT-0	Category 0
CCC	Common Control Channel
CCDF	Complementary Cumulative Distribution Function
CD	Code Domain
CDMA	Code Division Multiple Access
CH	Cluster Head
CN	Core Network
CR	Cognitive Radio
CRN	Cognitive Radio Network
CSI	Channel State Information
CSMA	Carrier-Sense Multiple Access
CSMA/CA	CSMA with Collision Avoidance
D2D	Device-to-Device
DACB	Dynamic Access Class Barring
DCMA	Dense Code Multiple Access

DQ	Distributed Queue
EAB	Extended Access Barring
EDT	Early Data Transmission
eMBB	enhanced Mobile Broad-Band
eNB	evolved Node B
FDD	Frequency Division Duplex
FDMA	Frequency Division Multiple Access
5G	Fifth Generation
H2H	Human-to-Human
HARQ	Hybrid Automatic Repeat Request
HTC	Human-Type Communications
IoT	Internet of Things
LDS	Low Density Spreading
LTE	Long Term Evolution
LTE-A	Long Term Evolution-Advanced
M2M	Machine-to-Machine
MAB	Multi-armed Bandit
MAC	Medium Access Control
MIMO	Multiple Input Multiple Output
mMTC	massive Machine-Type Communications
MPA	Message Passing Algorithm
MPR	Multiple Packet Reception
MTC	Machine-Type Communications

MTD Machine-Type communications Devices

MUD Multiple Users Detection

MUSA Muti-User Shared Access

MUST Multi-User Superposition Transmission

NB-IoT Narrow-band Internet of Things

nCH non-Cluster Head

NOMA Non-Orthogonal Multiple Access

NR New Radio

OFDMA Orthogonal Frequency Division Multiple Access

OMA Orthogonal Multiple Access

OSA Opportunistic Splitting Algorithm

PD Power Domain

PDSCH Physical Downlink Shared Channel

PRACH Physical Random Access Channel

PSD Power Spectral Density

PSM Power Saving Mode

PUSCH Physical Uplink Shared Channel

QoS Quality of Service

RAN Radio Access Network

RA Random Access

RACH Random Access Channel

RAR Random Access Response

RB Resource Block

RF	Radio Frequency
RL	Reinforcement Learning
RSS	Received Signal Strength
SCMA	Sparse Code Multiple Access
SIC	Successive Interference Cancellation
SJD	Successive Joint Decoding
SINR	Signal-to-Interference-plus-Noise Ratio
SNR	Signal-to-Noise Ratio
SR	Scheduling Request
3GPP	Third Generation Partnership Project
TDMA	Time Division Multiple Access
TR	Traffic Regulation
TST	Total Service Time
UCB	Upper Confidence Bound
UE	User Equipment
URLLC	Ultra-Reliable Low-Latency Communications
WUS	Wake-Up Signal

1. Introduction

1.1 Internet of Things and Machine Type Communications

The upcoming era of wireless communications is going to evolve in an entirely different way, thanks to the Internet of Things (IoT) that allows the communications of everyday objects. As the enabling technology for IoT, machine-type communications (MTC) is an essential target scenario for 5G networks and beyond [1]. MTC is about the automated data transmission from different devices with little or no human intervention. The applications of IoT and MTC are being extensively deployed in different domains including smart cities, industrial automation, eHealth, among others. The heterogeneity of the supported applications forces diverse quality-of-service (QoS) requirements. Based on that, MTC is generally categorized into massive MTC (mMTC) and ultra-reliable low-latency communications (URLLC) that target massive and mission critical IoT applications, respectively.

The diverse requirements of these applications have pushed different standardization bodies, such as Third Generation Partnership Project (3GPP), towards specifying the requirements and challenges of MTC [2, 3]. Indeed, cellular networks, such as Long Term Evolution-Advanced (LTE-A), are envisioned to be an excellent candidate to support MTC as these networks are mature, well-established networks that ease fast and low-cost deployment of MTC [4]. Additionally, cellular networks provide ubiquitous coverage, reliability, and controlled interference environment over the licensed band. However, these networks are well-suited to human-type communications (HTC), and have to be adapted to the different MTC features and requirements. As specified by the 3GPP [3], unlike the high power, costly HTC devices, MTC devices are inexpensive, low power, and usually low mobility devices. Furthermore, MTC devices are characterized with infrequent, small amount of data, and mainly uplink traffic. Mostly important, the number of MTC devices is expected to highly surpass that of the HTC devices.

The tremendous number of MTC devices that are typically wireless sensors, actuators, or

smart meters, connected with small amount of data, calls for urgent optimization for the existing wireless networks to be able to support this new communication paradigm. Specifically, the existing networks have to comply with the new requirements of MTC allowing low-power, low-rate and reduced complexity technologies. In this regard, the 3GPP has exerted considerable efforts by introducing new low-cost user equipment (UE) categories with reduced capabilities [5]. On the other hand, innovative communication protocols that are capable of handling MTC-related challenges such as energy efficiency, scalability, latency, coverage and security, are needed for the efficient deployment of MTC. From the early stages, the 3GPP has identified network congestion, also referred to as overload control, as first priority area of improvement [6]. Additionally, signaling overhead has been realized among the major challenges for MTC services.

Due to the explosion in the number of connections expected by MTC, network congestion is likely to happen at different levels in the network including radio access network (RAN), core network, and signaling network [1]. Out of these, RAN congestion problem is crucial due to the limited available resources at the access side. Without efficient access control, when massive number of MTC devices attempting to access the network concurrently, or even in a short period of time, RAN congestion occurs. This situation could lead to severe access delays, packet loss, or even service unavailability. The legacy coordinated access technique for cellular networks was shown to be inefficient for MTC due to the large signaling overhead and high latency [7]. Consequently, the need for efficient mechanisms to guarantee network availability and assist in meeting the performance requirements under such MTC load becomes inevitable.

1.2 Motivation

The key objectives of this thesis are to:

1. Study the challenges of supporting MTC as an integral part of the promising IoT for future 5G wireless networks and beyond;

2. Provide practical distributed and scalable mechanisms that are capable of efficiently tackling the MTC access problems including massive connectivity and QoS provisioning;

3. Develop analytical and simulation frameworks for the proposed techniques.

The proposed work in this thesis is important as it addresses an important and timely topic.

In particular, our proposed work is related to providing efficient uplink access mechanisms to MTC with a focus on handling network congestion while enhancing the resources utilization and the devices' QoS. As indicated before, MTC is considered a major backbone for the future IoT. The current aspiration for smart cities, smart homes, intelligent transportation systems, wearable, among others, relies on the effective deployment of IoT technologies.

The 3GPP has identified overload control as a key area of research [6]. The conventional random access (RA) mechanism used in Long Term Evolution (LTE) networks is not suitable for MTC. With MTC massive access, this mechanism is rendered a performance bottleneck due to the increased collisions that leads to severe congestion. Particularly, with the limited number of available preambles in existing networks, several MTC devices may concurrently select the same preambles, resulting in tremendous collisions and wastage of resources. In addition, the large signaling overhead needed to perform the RA mechanism is highly expensive for MTC devices with short packets. The 3GPP recommended a number of techniques that applied modifications to this RA scheme such as access class barring (ACB), pull-based and backoff schemes [6]. The major idea of these schemes is to limit the number of contending devices. Other works in literature were proposed such as dynamic ACB (DACB) [8], distributed queuing approach [9], among others [10]. However, the resulted performance falls beyond the great expectations from MTC deployment. In fact, the collisions could be reduced to some extent but the access delay remains relatively high especially for dense networks. Hence, there is a need for more practical solutions that respects the specific requirements of MTC considering both latency and scalability.

Recently, non-orthogonal multiple access (NOMA) has received great attention as a promising enabling technique for beyond 5G wireless networks [11, 12]. It allows multiple users to non-orthogonally share the same resource by multiplexing them either in power or code domain. From the information theoretic perspective, orthogonal multiple access (OMA) is strictly suboptimal in multi-user systems [13]. Thus, using NOMA with MTC can boost the spectral efficiency while handling the massive connectivity and reducing latency. These promising gains come at the expense of a more complex receiver that is able to decode the superposed signals. For uplink NOMA, successive interference cancellation (SIC) decoder exists at the base station (BS) which is fortunately acceptable for low-budget MTDs with mostly-uplink traffic.

Over the past few years, several variants of the RA scheme of cellular networks were introduced to adapt it for MTC access. However, without discounting the efforts of optimizing the

RA scheme, due to the increased collisions and prohibitive overhead, it falls short to support MTC with reduced latency and guaranteed reliability [7]. Alternatively different uplink access techniques such as grant-free [14], and fast uplink grant [15] are found promising in tackling massive connectivity while avoiding the shortcomings of the conventional RA. Particularly, grnat free schemes allows the MTDs to transmit directly to the BS without sending scheduling requests. On the other hand, using fast uplink grant, the BS allocates the resources directly to the MTDs without receiving a scheduling request. Utilizing NOMA with these access mechanisms can significantly boost the access performance of MTC.

Generally, efficient NOMA requires the design of advanced pairing and power allocation techniques, as well as powerful channel coding schemes to ease the decoding task [16]. Usually, this is done with the help of the BS which has full information about the users. However, for mMTC gathering such information from massive number of MTDs is impractical, specifically with grant-free access which is uncoordinated by nature. This poses the challenge of the efficient design of the contention units used by the MTDs for transmissions, in addition to the ability to resolve the collisions among the multiplexed MTDs. In fast-uplink grant, NOMA is promising such that the grant is assigned to multiple MTDs simultaneously. Consequently, the probability of wasting the resources assigned to inactive MTDs could be decreased. However, distributed low-complexity NOMA techniques are also needed to avoid any excessive overhead for information exchange between the MTDs and the BS. Keeping in mind the complexity of the NOMA receivers, it is important to control the number of MTDs sharing the same resource to strike balance between system performance and complexity. It is also worth pointing out that in such environments of uncertainties about the devices activity and/or parameters (e.g. QoS requirements), machine learning tools are promising to boost the overall performance.

1.3 Thesis Contributions

The contributions of this thesis can be summarized as follows.

- We address the network congestion problem of MTC. Assuming the devices access the network in a highly synchronized manner, we adopt a Beta-distributed traffic to model congestion as specified by the 3GPP. We propose a distributed overload control algorithm that is able to efficiently allocate the available network resources to massive number of devices within bounded delay and with reduced overhead [17]. Additionally, the algorithm

overcomes the resource wastage problem resulted from RA collisions, while providing more access opportunities. Thus, it achieves full resource utilization and reduced access delay. Furthermore, we study some cases of non-idealities that could occur in real situations and propose possible solutions for each case [18]. Additionally, we propose an updated version of the algorithm that includes a traffic regulation mechanism which effectively reduces the energy consumption of the devices.

- We assess the performance of the proposed overload control technique considering that the devices have either finite or infinite retransmission limits. We also provided a method for the estimation of the number of backlogged devices in the network which is neither known by the devices nor the BS. Furthermore, we provided an analytical framework that evaluates the efficiency of the proposed algorithm in terms of access delay, energy consumption, and blocking probability. For the sake of comparison, the problem is formulated as a binary integer programming (BIP) where the performance of the proposed practical technique is compared to the optimal one [19].

- According to recent studies that shed light on fast uplink grant as a promising alternative for the legacy RA technique, we adopt it in our work [20]. Particularly, we propose a two-stage NOMA-based fast uplink grant technique to schedule the devices without the need for sending access requests [21]. Fast uplink grant is known to have two major challenges, namely active set prediction and optimal scheduling. In our proposed work, we employ multi-armed bandit (MAB) learning to schedule the devices with no information at the BS about their QoS requirements or channel state information (CSI). In addition, a two-user NOMA scheme is proposed to limit the wastage of the resources due to source traffic prediction errors, whereas pairing is done in a distributed manner. Hence, a proper decoupling of the challenges associated to fast grant is attained as well as overhead reduction. We devise an abstraction model for the source traffic prediction. Consequently, the performance of the proposed scheme is analyzed considering average resource wastage, and outage probability [22]. A BIP formulation is also provided to show the potential of the two-stage algorithm in enhancing the access performance of MTC with enhanced pairing and power allocation.

1.4 Thesis Organization

The rest of the thesis is organized as follows:

In Chapter 2, we provide a relevant background for the topics of this thesis. We begin by giving an overview of IoT and machine-to-machine (M2M) communications including potential supporting technologies, main features and associated challenges for efficient deployment of IoT services. We also briefly illustrate the standardization efforts exerted by the 3GPP for supporting cellular MTC. Then, we survey the different uplink access techniques employed for MTC whereas the challenges and some of the state-of-the-art proposals of each technique are illustrated and compared. Then, we present the basics of NOMA in both uplink and downlink directions while emphasizing its potential gains as compared to orthogonal multiple access (OMA). Additionally, we give an overview of the applications of NOMA in both power and code domains. Moreover, we depict the promising gains of employing NOMA with MTC. Afterwards, we review the opportunistic splitting algorithm (OSA) as it is adopted in our work. Finally, we present the benefits of machine learning in supporting MTC, followed by a more in-depth description of MAB learning.

In Chapter 3, we address the massive access of MTC and the consequent congestion and resource wastage problems. We provide a literature review of the works studied the same problem. Next, the system model and assumptions are provided. Then, the procedure of the proposed access algorithm, referred to as M2M-OSA, is described. Next, we provide a method for the estimation of the number of backlogged users in the network that is important for the efficient operation of the algorithm. Afterwards, a number of non-ideal situation are discussed along with possible solutions. The performance evaluation and simulations results for the proposed algorithm as compared to one of the state-of-the-art proposals are provided.

In Chapter 4, we provide an analytical framework for M2M-OSA considering the activation of the devices follows Beta distribution. Particularly, we provide analytical expressions for different performance metrics including access delay, total service time, and energy consumption for the cases of infinite and finite retransmission limits. In the finite retransmission case, the blockage of the devices becomes possible. Hence, the analytical expression for the blocking probability of the M2M-OSA is derived. Afterwards, the BIP formulation of the studied problem is provided. We also present a modified version of the proposed M2M-OSA that allows for traffic regulation and energy saving. Finally, the simulation results that verify the accuracy of

the analytical expressions are depicted.

In Chapter 5, we adopt the fast uplink grant access technique as an access mechanism for MTC. Particularly, we utilize NOMA to better handle the associated challenges of fast uplink grant. We begin by a literature review of the related works. Then, the system model and assumptions are presented. Also, we provide an abstraction model for the source traffic predictor to evaluate the prediction error. Next, the problem statement is presented and the proposed two-stage access scheme is described. Due to massive access, the optimal pairing for the devices is known to be very complex. Hence, a quasi-optimal formulation is adopted that is based on BIP for the sake of comparison. The simulation results that show the effectiveness of the proposed scheme are illustrated.

In Chapter 6, we analytically study the major challenge of fast uplink grant techniques related to the resources wastage considering both OMA and the proposed NOMA scheme. In this regard, a modified abstraction error model for the source traffic predictor is introduced considering both false-alarm and miss-detection probabilities. Based on that we theoretically assess the performance of the proposed NOMA-based fast grant technique in terms of resource wastage and outage probabilities. In addition, the decoding order and the activity probability of the MTDs were carefully considered. Finally, a simulation framework is provided to evaluate the proposed scheme performance and verify the accuracy of the analytical expressions.

In Chapter 7, we present a brief summery of our investigation and some important conclusions. We also suggest some potential topics for future research.

2. Background and Literature Review

In this chapter, a brief background pertaining to the topics of this thesis is provided, including IoT and M2M communications along with their main applications, specifications and challenges. We also give an overview of the standardization efforts done by the 3GPP for MTC support, and the various uplink access techniques adopted for MTC and the associated challenges. In addition, we discuss how to utilize the promising NOMA scheme for MTC as well as its potential for efficient massive access support. Finally, we review some of the algorithms and machine learning tools that are utilized in the work proposed in this thesis.

2.1 Internet of Things (IoT) and Machine-to-Machine (M2M) Communications

M2M communications refer to an automated data transmission among devices without or with minimal human intervention. Particularly, the devices are equipped with sensing, computing, and communication capabilities to interact with the surrounding environment, collect data and take actions which forms the so-called Internet of things (IoT). The major objective of IoT is to enhance the quality of everyday life of the end-user community which has a significant impact towards the growing of the world's economy [23]. The applications of IoT and M2M communications are being extensively deployed in different domains including smart cities, smart homes, industrial automation, intelligent transportation systems, and eHealth, among others [23]. An example of an IoT network with different applications is shown in Figure 2.1.

M2M communication is application-oriented which means that not all the applications have the same characteristics. However, a number of features that can be individually activated according to the supported service have been identified such as [3]:

- **Infrequent, small data bursts:** Usually M2M devices generate infrequent and small-sized data packets. This feature imposes a limitation on the amount of traffic overhead.

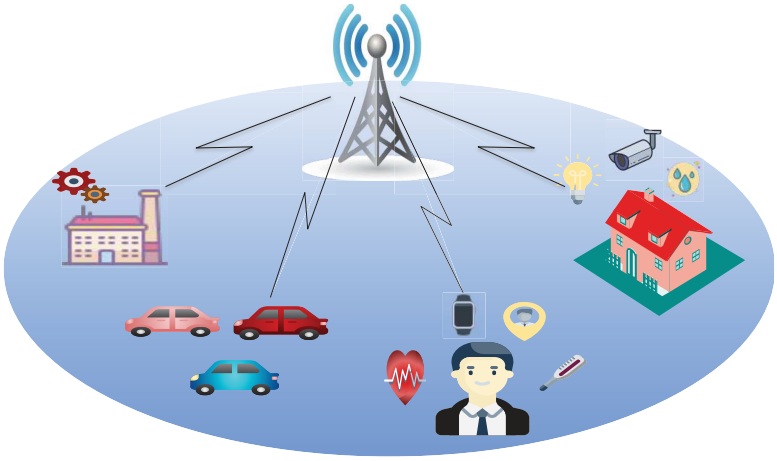

Figure 2.1: An example of an IoT network supports different applications such as industrial automation, intelligent transportation systems, eHealth and smart homes

Particularly, the connection establishment required for the current LongTerm Evolution (LTE) systems is considered expensive for M2M with small data transmissions. Data aggregation could be a possible solution for this issue.

- **Low mobility:** Unlike human-to-human (H2H) services, M2M communication should support low mobility devices. A number of potential M2M applications such as metering devices, weather monitoring sensors, and surveillance devices are known to be with low mobility [24].

- **Massive deployment:** Several M2M applications incorporate huge number of devices. Additionally, the diverse applications of M2M communications would lead to an explosion in the number of connected devices. Thus, the system should be able to efficiently support network scalability, otherwise network congestion that results in severe performance degradation occurs.

- **Low cost:** For mass-scale deployment of M2M devices, the devices should be of low cost. The low-cost feature results in a degraded transceiver performance and limited devices' computation capabilities [25]. In Release 12, the 3GPP has defined a new profile, referred to as category 0 (CAT-0), for low-cost devices [26].

- **Low power:** MTC devices are commonly energy-constrained. Therefore, guaranteeing low power consumption is essential for the operation of M2M services. For instance, metering applications are with large number of devices deployed in scattered areas where

the regular replacement of batteries is not always possible. Hence, long battery lifetime is highly desired.

- **Reliability and security:** These features must be met to ensure robust communications over M2M networks. Especially, M2M devices could be deployed in unprotected areas where the risk of malicious attacks increases. Additionally, the limited computational capabilities of the low-cost devices may be an obstacle to perform existing security schemes [27].

- **Very diverse quality of service (QoS):** The heterogeneity of M2M applications forces diverse QoS requirements. Hence, for proper M2M operation, different QoS requirements should be satisfied.

- **Time controlled:** M2M devices with time-controlled feature can only transmit or receive data in a predefined time intervals. Outside the access grant time interval, the network operator shall reject the access requests of M2M devices, or allow them with different charges.

- **Time tolerant:** The network operator can restrict or defer the access attempts of M2M devices with this feature in some cases such as network congestion.

- **Priority alarm:** For some applications, M2M nodes may need priority access over other nodes especially in case of emergency.

Due to the diverse applications and heterogeneous requirements of M2M communications, 5G is to support three main categories: enhanced mobile broad-band (eMBB), massive machine-type communications (mMTC), and ultra-reliable and low-latency communications (URLLC) [28]. Except for eMBB, the other use cases are related to M2M communications which reflects that the standardization of M2M has become inevitable. Particularly, mMTC target massive IoT applications where enormous number of devices with relaxed latency constraints are deployed. For instance, residents of smart homes can deploy sensors to automatically open their garage upon arrival to their homes, switch on lights, and control other smart appliances. In contrast, URLLC target mission-critical IoT applications with strict latency and reliability requirements. Usually, mission-critical applications are related to health, traffic, and industrial domains where failures could result in severe consequences. Table 2.1 shows some M2M applications associated with their main requirements [27].

Table 2.1: Examples of MTC applications and their requirements

Service Functions	Application Examples	Main Requirements
Metering	Electric power, gas, and water	Support of a massive number of MTC devices with small data bursts and high coverage
Control systems and monitoring	Industrial automation, and real-time control	High mobility and low-latency data transmissions
Tracking	Fleet management and asset tracking	High mobility and low power consumption
Payment	Point of sale and vending machines	High level of security
Security and Public safety	Surveillance systems, home security, and access control	High reliability, high security, and low latency

2.2 Cellular Networks for Machine-Type Communications

The 3GPP has identified two types of communications related to M2M communications: communications between the M2M devices themselves and the communications between the M2M devices and the server [27]. In this section, we are concerned by the latter case that can be provided by cellular networks. In the context of LTE/LTE-A, M2M is also referred to as Machine-Type Communications (MTC). In this regard, cellular networks are known as an excellent candidate for the support of MTC. Particularly, cellular networks are mature, well-established networks that enable fast and low-cost deployment of MTC [1]. Additionally, cellular networks provide ubiquitous coverage that ease the widespread deployment of MTC with mobility support. Furthermore, the controlled interference environment over the licensed band provides communication links that are more reliable. However, cellular networks were designed to support high-end H2H devices with powerful capabilities, large capacity batteries and costly hardware. This is completely different from the MTC features mentioned in Section 2.1 where the devices are inexpensive with low-power. Moreover, MTC is characterized by infrequent, small-sized data packets with traffic mostly in the uplink direction. In contrast, the legacy human-type communications (HTC) traffic is characterized by bursts of data with higher demand in the downlink [27, 29]. Importantly, the number of MTC devices is very huge, orders of magnitudes of the HTC devices. These differences between HTC and MTC bring forward a number of challenges that should be efficiently tackled such that MTC can be genuinely supported. In the following we illustrate a number of the major challenges of MTC.

2.2.1 MTC Challenges

- **Energy Management:** One of the major challenges of M2M communication is reducing the power consumption especially for set-and-forget M2M applications where it is almost impossible to replace or recharge batteries. One of these applications is the animal tracking MTC devices in natural world with high mobility [3]. In addition to battery-constrained devices, efficient energy management is essential for both economic and environmental reasons [30]. Therefore, operating at extra-low power mode is required. Mostly, the power is consumed in radio transmissions and channel access. A number of suggested solutions can be employed to reduce the devices' energy consumption such as modifying signaling and medium access control (MAC) protocols to reduce collisions during channel access, sleep scheduling, and power control. Furthermore, aggregation of control information as well as support of device-to-device (D2D) links are possible solutions [27].

- **Enhanced coverage:** Indeed, the wide deployment of MTC requires coverage enhancements over the current LTE cell footprint. This is due to the reduced capabilities of MTC devices, in addition to the fact that several MTC applications require the deployment of the devices in challenging locations such as basements or indoor areas. The 3GPP tackled the standardization of MTC towards enhanced coverage in Release-13. In order to accommodate the new MTC requirements, the CAT-M UE category is added in Release-13. Particularly, CAT-M features a coverage enhancement of more than 15 dB, enabling the reach of the devices in locations with excessive penetration losses [26]. Different coverage enhancement techniques for different physical channels were proposed. For instance, repetition across time which allows for the accumulation of the energy to increase the signal strength and hence compensates for penetration losses can be used. Also, frequency hopping is utilized to compensate for the lack of frequency diversity of narrow-band UEs. Other techniques can also be employed such as power spectral density (PSD) boosting, joint channel estimation, and retransmission using hybrid automatic repeat request (HARQ), among others [27]. To further achieve better coverage enhancements, several techniques can be jointly utilized.

- **Cost reduction:** In the current HTC devices, the communication unit (i.e. the cellular modem) represents only small part of the device's cost which renders cost-reduction less

important [31]. In contrast, the cost of the communication unit for MTC has to be significantly decreased. Consequently, the widespread integration of MTC devices within other types of devices such as wearable devices and utility meters would be possible. In this regard, the 3GPP considered various cost-reduction techniques including reducing the data rates and the devices' bandwidth, supporting single antenna, and half duplex operations [31].

- **Congestion control:** With the emerging M2M applications, it is expected that a numerous number of devices are deployed in a specific area. Network congestion is likely to occur especially in the radio access network (RAN) due to mass concurrent data and signaling transmissions. This may result in intolerable delays, packet loss or even service interruption. Hence, the 3GPP has identified RAN congestion control as the first priority improvement area [6]. The 3GPP also recommended a number of candidate solution to protect the RAN from PRACH overload such as access class barring (ACB) schemes, separate PRACH resources for MTC, and back-off mechanism. A review of the 3GPP overload control solutions is provided later.

- **QoS provisioning:** The heterogeneity of QoS requirements of different M2M applications poses a major challenge for MTC to operate effectively. Particularly, the low-power feature desired for MTC devices for energy saving limits the wireless communication performance. Moreover, the huge number of devices in M2M networks complicates the resource allocation process to be able to ensure QoS and reliability requirements of different devices. On the other hand, the mechanisms employed for M2M communications should be simple and yet effective to be capable of attaining the QoS requirements whereas at the same time preserve the low-complexity requirement for MTC devices. These all make guaranteeing the QoS for M2M applications a challenging issue. Efficient mode selection methods that allow the MTC devices to select either to communicate directly with the eNB or via a gateway help in guaranteeing QoS [32]. Additionally, grouping based on QoS requirements can alleviate the congestion in the PRACH and hence achieve lower delays. A QoS-aware random access transmission was proposed in [33].

- **Efficient group management:** Grouping is a vital mechanism to handle M2M communications especially with massive access. However, in the context of M2M, grouping becomes more challenging. For instance, the selection of the group coordinator and man-

aging the groups are more challenging with large number of groups or large number of members per group. Furthermore, the formation of the groups should be done adaptively according to the traffic conditions. Also, peer-to-peer communications between group members should be provisioned [32].

In addition to the above mentioned challenges, the cellular operators and service providers have a natural concern related to the impact of MTC on the HTC services [34]. It is extremely important to provide MTC services while minimizing the impact on the existing voice and data services. The impact of MTC on HTC can be mitigated by means of scheduling prioritization, time controlled access for M2M such that some slots are forbidden for MTC access, and semi-persistent scheduling to reduce overhead [35]. In [34], with a focus on the random access (RA) procedure of LTE/LTE-A, two priority-based access schemes were introduced. It was shown that MTC is slightly affected by the priority methods. So, network operators should always give priority to HTC. For delay-constrained MTC devices, it is recommended to form groups of these devices that have similar priority as HTC devices. Hence, they are minimally impacted by congestion problems. In [36], a complete framework that studies the coexistence of HTC and MTC considering both random access and uplink scheduling is provided. In the following, we shed light on the standardization efforts exerted for cellular support of M2M communications.

2.2.2 MTC Standardization

Cellular networks are promising for supporting IoT services including MTC. However, as indicated earlier, major differences between M2M and H2H users exist that make the adaptation of the current cellular networks to M2M communications more challenging. Particularly, LTE networks should continue the evolution towards more efficient mobile broadband platforms supporting high-end devices while at the same time providing unified reliable and cost-effective connectivity for MTC services [37]. To that end, the 3GPP started the standardization of MTC as early as Release-10 and continued the work through future releases [38]. In the following, we provide examples of the tasks tackled by different releases.

- **Release-10**
  - Enhancements on the core network (CN) to enable M2M
  - Overload and congestion control on both CN and random access network (RAN)

- **Release-11**
  - UE power preference indication was introduced to help in reducing power consumption by allowing the device to indicate to the network that it prefers the operation at low power. The network then configures the parameters adaptively to allow, for instance, for longer sleep time and reduced measurements.
  - RAN overload control
  - Refining existing requirements, and use cases
- **Release-12**
  - Defining new low-cost UE category, namely category 0 (CAT-0) for MTC devices with the following physical constraints [39]:
    * A single RF chain and baseband processor
    * Half duplex operation for FDD
    * Reduced peak rate (e.g. 1 Mb/s, without multiple-input multiple-output (MIMO), or mobility) in downlink and uplink
  - Introducing the power saving mode (PSM) where the UE remains registered to the network but not reachable for downlink traffic
- **Release-13**
  - Further complexity reduction on top of CAT-0 by introducing UE category-M1 (CAT-M) [40]
    * Reduced RF bandwidth from 20 MHz to 1.4 MHz in both downlink and uplink
    * Reduced transmission power to allow for on-chip power amplifier
  - Power saving and coverage enhancements
  - Narrow-band IoT (NB-IoT) or CAT-N was introduced [26, 31]
    * Supporting the transmission of small data amounts with extended coverage
    * Maximum channel bandwidth is reduced to 1.08 MHz, or 6 LTE resource blocks (RBs)
    * Enhancing the link budget (20 dB improvement over conventional LTE-A)

  - The requirements were attained by employing different techniques, including repetition techniques for control channels, frequency hopping, and uplink power spectral density (PSD) boosting [41].

- **Release-14**
  - Last release before entering the era of 5G [5]
  - Introducing new UE categories with even lower power consumption (e.g, Category-M2) [5, 42]
  - Enhancing the ability for LTE-M (i.e. device category M1, and category M2) to support voice services
  - Introducing the multicast feature for LTE-M devices
  - Positioning enhancements and mobility improvements for enhanced connectivity
  - Introducing higher data-rate feature to support some MTC use cases such as healthcare, and surveillance cameras [42]
  - Investigating various latency reduction schemes by means of [43, 44]:
    * Increasing the frequency of Semi-Persistent Scheduling (SPS)
    * Reducing transmission and processing delay (shortening the transmission time interval (TTI))
    * Reducing random access channel (RACH) procedure delay during handover
    * Fast uplink grant instead of scheduling requests (SR)-based access

- **Release-15**
  - First release to introduce 5G [5]
  - Advancements in improving the power efficiency by reducing downlink channel monitoring by the UEs in idle mode (Wake-up signal (WUS) feature)
  - Introducing new use cases including wearables devices, and support for standalone operation
  - Allowing for early data transmission (EDT) such that the data is transmitted at the beginning of the connection setup procedure for further latency reduction [45]
  - Introduction of 5G new radio (NR)

- **Release-16**
  - The 3GPP introduced additional enhancements for both MTC and NB-IoT [46]
  - Enhanced spectrum efficiency and reduced energy consumption for mMTC by:
    * Enhanced early data transmission and UE-group wake-up signaling
    * Reducing the control signaling overhead by multi-transport-block scheduling in both the downlink and uplink
    * Employing preconfigured resources in idle mode for uplink transmission, allowing the device to avoid time-consuming RA process
    * Improving the performance in terms of NR/LTE-MTC co-existence
  - Providing higher efficiency in case of LTE-MTC standalone operation for cells only providing LTE-MTC access

To sum up, significant efforts are being exerted in the standardization process of MTC. As seen, both NB-IoT and LTE-MTC were introduced in Release-13 to provide support for low-complexity, deep coverage, mass deployments of LTE IoT devices. Additional IoT support was added in Release-14, such as positioning and multicast, and data rates to suit a broader range of usages. Release-15 and Release-16 have enhanced the two technologies further by introducing features that allow reduced signaling and control channel overhead for transmitting messages, extensions to battery life, channel quality reporting to the network, and coexistence with NR. In Release-17, it is proposed to add or enhance features which support the broadening use cases for cellular IoT, address lessons drawn from deployments and trials, and support the long-term life cycle of the two technologies [47].

It is worth mentioning that other non-cellular technologies either short-range or long-range can also be utilized for the support of MTC. For instance, WiFi, Zigbee, and Bluetooth Low Energy (BLE) are among the candidate short-range technologies [48]. These technologies are cheap alternative with the ability to support low-power transmissions [27]. However, low rate, limited coverage, lack of efficient backhaul, and severe interference due to operation over unlicensed frequency bands are major obstacles that limit their widespread support for MTC [29]. For long-range technologies, LoRa and SigFox are among the strong candidates for non-cellular MTC support [28, 48]. A comprehensive survey of different technologies to support MTC can be found in [48, 49].

2.3 Overview of Uplink Access Techniques for MTC

A major challenge for MTC support in cellular networks, is the conventional random access (RA) mechanism of LTE networks. With massive number of connected MTC devices (MTDs), an immense traffic is generated that leads to network overload or congestion. This situation severely affects the QoS of the MTDs such that intolerable access delays, packet loss, or even service interruption are encountered. Mainly, for efficient support of MTC access the developed medium access control (MAC) protocols should satisfy certain requirements:

- Achieving high access success rate (throughput) to accommodate massive MTDs within limited resources;
- Provisioning of QoS with a focus on low-latency and reliability requirements for certain applications such as eHeath;
- Scalability with low energy consumption by means of controlling collisions;
- Implementing cost-effective design where efficient protocols are using simple, low-cost hardware.

In this section, we give an overview of the conventional RA scheme of LTE, along with the associated challenges of adopting it for mMTC. Moreover, we illustrate the proposed state-of-the-art modifications applied to it to be adapted for mMTC. We discuss the shortcomings of these proposals pointing out that with the new features of MTC the shift to new access techniques is vital. In this regard, we discuss both grant-free access and fast uplink grant access mechanisms. Following the same way, the associated challenges and the state-of-the-art proposals studied the new schemes are reviewed and compared.

2.3.1 Grant-Based Random Access

2.3.1.1 Overview of the RA Process of LTE/LTE-A

In LTE/LTE-A networks, the random access (RA) procedure has two forms, contention-free and contention-based. In contention free, the network assigns specific preambles to the user for transmission. Usually, contention-free is used for access requests that require high probability of success such as handover process. However, this procedure limits the scalability of the

network [50]. For MTC, the focus is on the contention-based RA to be able to serve larger number of devices. Generally, the RA procedure is used (i) for initial access request (ii) after link failure (iii) to acquire uplink synchronization and (iv) during handover to another evolved Node B (eNB). This RA procedure is allowed only during specific periods called RA slots. It also has dedicated uplink resources that are divided between two channels namely, physical random access channel (PRACH), and physical uplink shared channel (PUSCH). Initially, the eNB broadcasts the preambles available for contention. Afterwards, the contention-based RA process, which is a 4-step handshaking between the user equipment (UE) and the eNB, proceeds as follows [51], [52]:

1. **RA preamble transmission:** each UE randomly selects one preamble and transmits it using the PRACH on the next RA slot.

2. **Random access response (RAR):** the eNB detects the active preambles and respond to each successfully decoded preamble by a RAR message via the physical downlink shared channel (PDSCH). The RAR message includes RA preamble ID, synchronization information, uplink resource grant, among others.

3. **Connection request:** the UEs that successfully received RAR messages can determine the allocated uplink resources (i.e. PUSCH) and use these resources for access request transmission. The failed UEs should reattempt later.

4. **Contention resolution:** the eNB acknowledges the reception of successfully decoded messages from step 3. Successful UEs then proceed onto data transmission. The collided UEs that do not receive the contention resolution message are failed. They should backoff and reattempt later.

During the first step, multiple UEs could select the same preamble and transmit it. In this case, collision occurs that is usually not detected by the eNB at the second step. Hence, the RAR message is sent, the collided UEs transmit at the same PUSCH, and the resource is wasted. In different scenarios, it could be assumed that the eNB is able to detect one of the collided UEs which is known as the capture effect [53]. Commonly, the capture effect is neglected and non of the collided UEs succeed [6]. Regarding the collision detection, some proposals assumed that the eNB can detect the preamble collision at step 2, and thus the RAR message will not be sent [6, 54]. However, this cannot be guaranteed, especially when the collided UEs are at

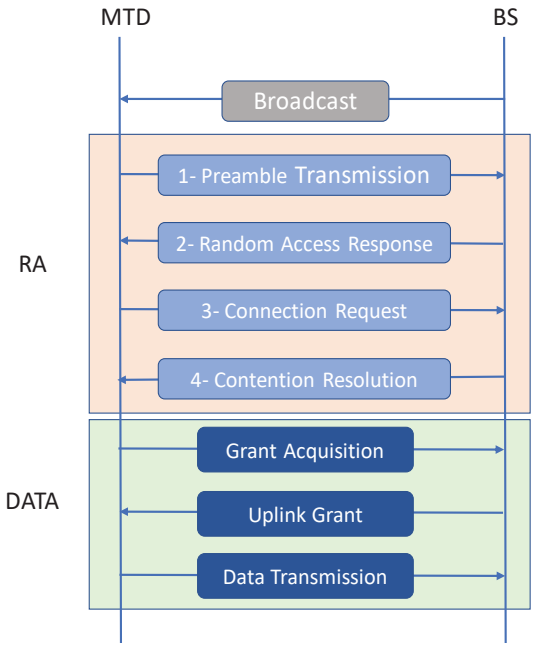

Figure 2.2: Grant-based RA process, after completing the 4-way handshaking of RA the data transmission phase starts.

the same distance from the eNB [55]. Hence, the eNB would transmit the RAR message and the UEs use the same PUSCH resource to transmit their requests. The collision is detected at step 3 where the uplink resources are wasted [34, 56]. Figure 2.2 illustrates the signaling of the grant-based RA process.

2.3.1.2 Challenges of RA with MTC

The efficient management of the RA process to support mMTC is known to be a challenging task. The RA resources are strictly limited as compared to the enormous number of MTC devices (MTDs) in massive networks. With bursty MTC traffic where large number of MTDs are activated simultaneously or within small period of time, network congestion and increased collisions are expected. Consequently, the MTDs suffer from excessive access delay, packet loss, and even service interruption. In addition, the large signaling overhead of the 4-way RA process is generally impractical especially with MTDs with small payloads. Both collisions and signaling overhead are sources of high energy consumption which is improper for energy-constrained MTDs [57]. It is worth mentioning that, the uplink resources reserved for the RA process are divided between the PRACH and the PUSCH. This further limits the success opportunities of the MTDs and could result in access blockage due to shortage in the PUSCH resources even if the preamble transmission was successful. Hence, the optimum division of

the RA resources is another challenge.

2.3.1.3 State-of-the-Art Proposals for RA with MTC

Formerly, the 3GPP identified this RA process as a key area of improvement for MTC support. In this regard, the 3GPP specified a number of candidate solutions to enhance the performance of PRACH in overload situations. The major idea behind these schemes is to limit the contention among the MTDs as follows [6]:

- **Access class barring (ACB) scheme:** This scheme depends on defining an access probability (AP) and access class (AC) barring time broadcasted by the eNB to the MTC devices. Then, each device before initiating a random access draws a random number from the interval $[0, 1]$, known as ACB factor. If the picked ACB factor is less than AP, it will proceed to the random access. Otherwise, the device is barred for the AC barring duration. The ACB scheme was then modified to allow for several access classes where different access probabilities could be assigned to different classes of MTC devices. Furthermore, an extended access barring (EAB) scheme has been proposed to allow the network to restrict the access of certain delay-tolerant devices that are configured for EAB in overload situations. Hence, no need to define any new access classes.

- **Separate RACH resources for MTC:** In this scheme, separate RACH resources are assigned to H2H UEs and M2M devices such that the performance of H2H UEs is not degraded. This separation could be done either in the preamble domain or by assigning different RA slots to H2H and M2M users.

- **Dynamic allocation of RACH resources:** In this scheme, the eNB dynamically allocates the RACH resources to the MTC devices according to the load conditions. For instance, the frequency of the RA slots per frame can be increased in overload situations. However, there is a trade-off between the number of PRACH resources and the resources available for data transmission. Hence, the allocation of the resources should be efficiently optimized to improve the overall performance.

- **Backoff scheme:** This scheme is used to delay the random access reattempts of the users. Usually, the backoff time of H2H UEs is less than that of MTC devices. This scheme can improve the congestion in low overload situation, but not as effective in heavy overload situations.

- **Slotted access:** In this scheme, the eNB broadcasts an RA-cycle which is an integer multiple of a radio frame [32]. Depending on the RA-cycle and the device ID, an MTC device determines its allocated RA slot. The MTC devices can only access the network during their dedicated slots. A collision occurs when the number of MTC devices in a cell is larger than the number of unique access slots such that multiple devices share the same slot for network access.

- **Pull-based scheme:** If an MTC server is aware of when the MTC devices have data to transmit, it can inform the eNB. Hence, the eNB sends paging messages to the MTC devices. Upon receiving a paging message, the MTC device initiates the random access procedure. The eNB can control the number of paged devices according to the overload situation, resources availability, and load conditions. However, this scheme incorporates extra control signaling.

Afterwards, several proposals were introduced to adapt the RA process such that MTC congestion is mitigated (e.g. [51]). The majority of the proposals adopt one of two approaches. The first is to regulate the number of contending MTDs according to the available resources. ACB techniques and MTDs grouping are examples of this approach [51]. For optimal ACB, the barring factor is dynamically adapted based on the load conditions. Variants of ACB that allow two-stage contention, or multiple-class ACB based on the MTDs' priority and QoS requirements were proposed. Also, grouping can deal with congestion and reduce the signaling overhead. Generally, the MTDs are grouped based on their QoS requirements, application type, or location.

The second approach is to adapt the RA resources according to load conditions. This can be done by changing the PRACH configuration to allow more RA slots, or spatial grouping such that the same preambles are used by MTDs that are far apart [58]. Moreover, the use of multiple-RAR techniques, and virtual preambles can increase the access success probability since the eNB becomes able to identify the collided MTDs and assign them different resources [59].

Other than mitigating the congestion, the 3GPP took also the direction of optimizing the RA procedure itself for the new radio (NR) to support different preamble formats (short/long). Moreover, to accommodate the small amount of MTC data, instead of transmitting access request on step 3 of the RA process, fast data transmission can be adopted [56]. While all the aforementioned improvements help the accommodation of MTC in cellular networks, these might not be enough. This is because MTC has totally different features and requirements from

HTC. This renders the lengthy RA process, originally designed for HTC, improper to be used with massive MTDs with sporadic traffic.

2.3.2 Grant-Free / Uncoordinated Access

Given that the aforementioned RA process is followed by a grant-based data transmission, the need for reducing the signaling overhead and delay becomes mandatory. In this regard, the 3GPP took stepwise proposals to uplink grant-free transmission for new radio (NR) [60]. Particularly, although RACH-based and uplink-grant based orthogonal multiple access (OMA) is considered as the baseline for NR, the 3GPP recommends the investigation of grant-free scheme whether RACH-based or RA-less such that the reduction of overhead and delay can be achieved. In this section, we reflect the focus of the literature on RACH-less grant-free scheme. For simplicity, we refer to it as grant-free.

2.3.2.1 Overview of RACH-less Grant Free

In grant-free schemes, the active MTDs transmit directly to the base station (BS) without sending a grant request or going through the RA process [14]. This is more suitable for mMTC with short packets where expensive RA signaling cannot be afforded. The transmitted packet may contain metadata in addition to the actual data [52]. The metadata (i.e. pilot) contains preamble to facilitate device-activity detection and channel estimation. While minimal overhead is attained, one major drawback is the collisions that occur when multiple MTDs select the same uplink resource. In few scenarios this can be handled relying on retransmissions or adjusting the amount of the resources [60]. Figure 2.3 depicts the signaling of grant-free access.

2.3.2.2 Challenges of Grant Free

With mMTC, the lack of coordination inherited in grant-free schemes results in tremendous amount of collisions due to the limited resources for massive number of MTDs. Hence, more sophisticated solutions rather than retransmissions are needed to handle the collisions. Blind detection of the received data at the BS is another major challenge to apply grant-free schemes [14]. The reason is that the BS has no prior knowledge about the set of active users and their respective channel conditions. Thus, grant-free scheme requires the BS to perform active user detection, channel estimation and data recovery either jointly or separately. Another challenge is the synchronization among the MTDs due to omitting the RA process.

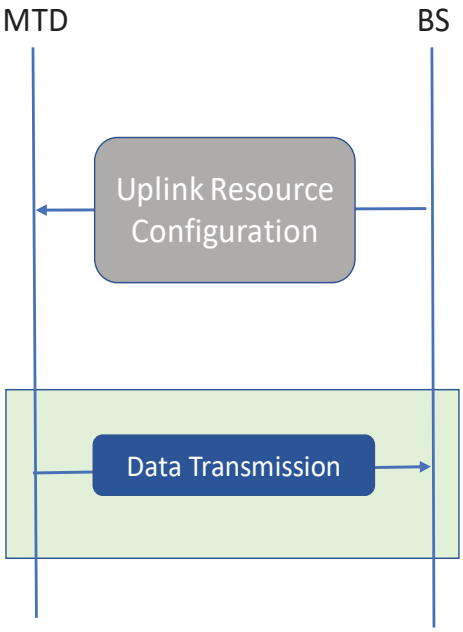

Figure 2.3: RACH-less grant-free access scheme where data is transmitted directly without grant, preambles and control may be combined with data.

2.3.2.3 State-of-the-Art Proposals for Grant Free

As indicated before collisions are a major obstacle such that grant-free scheme cannot stand alone to support mMTC. In literature, two approaches have been adopted to solve this problem. The first is to apply non-orthogonal multiple access (NOMA) for MTDs to share the limited resources with the cost of an increased receiver complexity. Usually NOMA is used with co-ordinated access where the BS has full information about the users which is clearly not the case with grant-free access. In [61], a distributed layered grant-free NOMA framework was proposed such that the cell is divided into layers with different power levels, and only devices of different layers could share the same resource without collisions. In this scheme, the MTDs can decide their own uplink channel and power level without the BS assistance. To further reduce collisions, enhanced access barring (EAB) was applied. A comprehensive survey of NOMA-assisted grant-free proposals can be found in [14]. The authors in [14] also refer to a number of potential solutions for the synchronization problem such as utilizing the downlink synchronization by the MTDs to adjust the uplink timing.

The second approach to overcome the increased collisions is to apply massive multiple-input multiple-output (MIMO) to exploit the spatial degrees of freedom in limiting the contention. In [52], the device activity detection problem of the grant-free scheme was studied with a focus

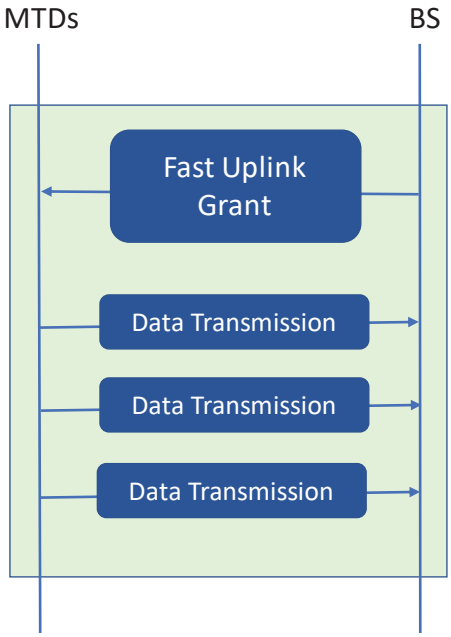

Figure 2.4: Fast uplink grant access where multiple MTDs are scheduled at once

on advanced compressed sensing techniques. The authors argued that massive MIMO can boost the device-activity detection accuracy remarkably for massive IoT connectivity. In grant-free access literature, it was noted that one common assumption is usually adopted to facilitate the tasks of active user identification, channel estimation and data detection. Although mMTC has massive number of MTDs, it is assumed that usually a small (unknown) subset of them is active at any given instant (e.g., see [52] and references therein).

2.3.3 Fast Uplink Grant

2.3.3.1 Overview of Fast Uplink Grant

In fast uplink grant, MTDs no longer need RA. Instead, the BS allocates the uplink grants proactively to the MTDs without receiving any scheduling requests [15]. Hence, both large signaling overhead and collisions are avoided. Particularly, with fast uplink grant the BS performs one level of signaling and the uplink grant for the entire system can be sent in one message which significantly decreases the signaling overhead. Moreover, by eliminating collisions, energy-constrained MTDs can save significant energy. Also, larger number of MTDs can be supported by utilizing the whole RA resources for uplink grant without dividing them for PRACH. Figure 2.4 shows the procedure of fast uplink grant scheme.

2.3.3.2 Challenges of Fast Uplink Grant

The major challenge of adopting fast grant is the selection of the scheduled MTDs [15]. The aforementioned promising gains of fast uplink grant come with the risk of wasting the granted resources if the scheduled MTD has no data for transmission (i.e. inactive). Additionally, a serious latency and reliability problem can occur if the MTDs wait for the grants more than their maximum tolerable delays. Consequently, the packets are dropped. This selection process requires the BS to predict the set of active MTDs at each time slot to avoid wasting the resources. Moreover, the granted resources should go to the optimal MTDs according to each network requirements. Specifically, the QoS requirements of the MTDs should be taken into consideration such that both latency and reliability can be guaranteed.

2.3.3.3 State-of-the-Art Proposals for Fast Uplink Grant

The employment of fast grant in tackling the access problems of mMTC is relatively recent. Thus, the state-of-the-art proposals studying this scheme are limited. In [15], the authors proposed a two-stage algorithm that shed light on some mathematical and machine learning tools that can be useful in source traffic prediction of the MTDs whether periodic or event-driven traffic. Moreover, they suggested to utilize multi-armed-bandit (MAB) learning or deep learning to solve the optimal scheduling of the MTDs.

With mMTC, to optimally schedule the MTDs, enormous information should be gathered by the BS related to the MTDs' QoS and channel state information (CSI) which is unrealistic. In a later work [62], MAB-based fast grant scheduling of MTDs was proposed. Particularly, the MTDs have rewards that depend on a weighted combination of the different QoS requirements. Therefore, the BS learns over time to select the MTDs such that the cumulative system reward is maximized. Although this technique gives promising results, it has a major problem that is its performance is dependent on the accuracy of the traffic prediction scheme. Knowing that the prediction becomes more sophisticated as the number of MTDs increases, it is encouraged to mitigate the effect of traffic prediction accuracy on the scheduling performance. Table 2.2 provides a comparative summary between all the reviewed uplink access schemes for MTC.

Table 2.2: Comparison between the reviewed access techniques for MTC

Constraint Forms	RA Grant-Based	Grant Free	Fast Uplink Grant
Motivations	• Compatible to LTE/LTE-A networks • Simple and distributed	• Minimal signaling overhead • Suitable for short-packet transmission	• Large signaling overhead and collisions avoidance • Single grant message for entire system
Limitations and Challenges	• Congestion with massive access • Increased collisions and high signaling overhead • Wastage of resources and energy • Division of limited uplink resources between PRACH and PUSCH	• Lack of coordination results in increased collisions • Blind data detection • Active MTDs identification • Channel estimation • Uplink synchronization	• Source traffic prediction • Optimal scheduled MTDs based on their QoS
Solution Approaches	• MTDs traffic regulation to match the resources • RA resources adaptation according to load conditions • optimizing RA process to support long/short preambles	• NOMA or massive MIMO for collisions • Compressed sensing for active user identification • ML and MTDs' signal sparsity for activity/data detection • Utilizing downlink synchronization for uplink	• ML techniques to limit information exchange
Common Assumptions	• Negligible capture effect • Collision detection at step 3 • Fast data transmission at step 3	• RACH-less • Small unknown portion of active MTDs (sparsity)	• the BS employs high-accuracy source traffic predictor
References	[6, 51, 58, 59]	[14, 52, 60, 61]	[15, 62]

2.4 Non-Orthogonal Multiple Access (NOMA)

In the past decades, orthogonal multiple access (OMA) schemes such as frequency-division multiple access (FDMA), time-division multiple access (TDMA), code-division multiple access (CDMA), and orthogonal frequency-division multiple access (OFDMA) have been employed for wireless networks [16]. In such OMA schemes, there is no interference among the users allocated the orthogonal resources. This facilitates the separation of the users' signals using rel-

atively low-complexity, cost-efficient detectors. However, OMA upper bounds the number of supported users by the available number of orthogonal resources. For multiuser systems, it has been theoretically proven that OMA cannot always attain the maximum sum rate [63]. In addition, the orthogonality among the users allocated different resources is vulnerable to distortion due to channel impairments [16]. Also, a single user served solely on each orthogonal resource block (RB) significantly affects the spectral efficiency. For instance, consider the case when it is inevitable to serve a user with poor channel conditions due to fairness or priority issues. In this case, the scarce resources allocated solely to this user is underutilized. Alternatively, allowing multiple users to concurrently share the orthogonal resources, known as non-orthogonal multiple access (NOMA), can boost both the spectrum utilization and system throughput.

NOMA is a promising technology for supporting the heterogeneous services of 5G wireless networks and beyond [11]. The key idea of NOMA is to support a number of users that exceeds the number of available orthogonal resources. Particularly, NOMA adds an extra dimension such that multiple users can transmit on the same frequency-time resource using different power levels or codes, referred to as power-domain (PD), and code-domain (CD) NOMA, respectively [11, 16]. Consequently, the spectrum utilization, network sum rate, and fairness among the users are all enhanced. However, these promising gains come at the expense of increased complexity of the receiver that employs sophisticated inter-user interference cancellation techniques to separate the superimposed signals. In the same context, the set of users that share the same orthogonal resource using NOMA is known as a NOMA cluster [28]. Thus, the complexity of the receiver increases with the increase of the size of the NOMA cluster (i.e. polynomial or exponential complexity increase) [16]. Joint decoding and successive interference cancellation (SIC) are among the well-known decoding techniques for NOMA systems [64].

2.4.1 Downlink NOMA

In downlink NOMA, two or more users are being served by the associated BS on the same orthogonal resource. In this case, the superposed signal transmitted by the BS to the i^{th} user of an N-sized NOMA cluster is formulated as [65]:

$$y_i = \sum_{n=1}^{N} h_i s_n x_n + w_i \tag{2.1}$$

where $n \in [1, N]$ is the user index within the NOMA cluster. h_i is the complex channel coefficient between the BS and user i, and x_n is the signal transmitted to user n such that $\mathrm{E}[|x_n|^2] = 1$. Also, s_n is a NOMA-dependent parameter that represents the power allocation, or the spreading sequence assigned to user n whether multiplexing is done in power or code domain, respectively. w_i is the noise at the receiver of user i including inter-cell interference with power spectral density $N_{0,i}$.

In [65], the authors showed the superiority of NOMA in achieving the multi-user capacity region as compared to OMA. Without loss of generality, the authors considered a 2-user PD-NOMA scenario where SIC is implemented at the receiver side. Hence, the user with better channel conditions employs SIC to eliminate the message intended for the other user before decoding its own. Based on that, the achievable rates of the two users are expressed as follows:

$$R_1 = \log_2(1 + \frac{P_1|h_1|^2}{N_{0,1}}) \tag{2.2}$$

$$R_2 = \log_2(1 + \frac{P_2|h_2|^2}{P_1|h_2|^2 + N_{0,2}}) \tag{2.3}$$

where user 1 is assumed to be the one with better channel conditions with P_1 allocated power. Also, P_2 is the power allocation for user 2 with poor channel conditions. Figure 2.5 from [65] depicts the capacity regions for the cases of symmetric and asymmetric channels, respectively. In symmetric channels, the signal-to-noise ratios (SNRs) of the two users are the same, whereas they are different in the asymmetric case. As seen, in the symmetric case the throughput region is identical for both OMA and NOMA. In contrast, the throughput region of NOMA with SIC is much wider than that for OMA in the asymmetric case. It shall be noted that the power split plays a significant role in the achievable rates of the users.

2.4.2 Uplink NOMA

In uplink NOMA, multiple users transmit concurrently on the same frequency resource to the BS. Hence, the received signal at the BS is expressed as:

$$y = \sum_{n=1}^{N} h_n s_n x_n + w \tag{2.4}$$

where $n \in [1, N]$ is the user index within the NOMA cluster of size N. h_n is the complex channel coefficient between the BS and user n, and x_n is the transmitted signal of user n such

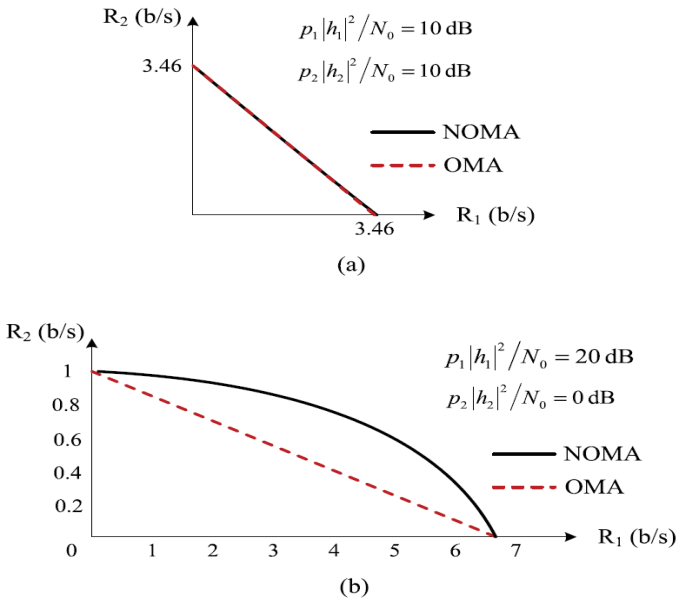

Figure 2.5: Throughput (capacity) region in downlink [65]

that $\mathrm{E}[|x_n|^2] = 1$. Also, s_n is a NOMA-dependent parameter that represents the transmission power, or the spreading sequence of user n whether multiplexing is done in power or code domain, respectively. w is the noise at the BS including inter-cell interference with power spectral density N_0.

From an information theoretic perspective, a comparison between the capacity regions of both OMA and NOMA for 2-user case with SIC was conducted in [65]. Applying PD-NOMA, the BS first decodes the signal having higher received power while treating the signal of the other user as interference. Then, eliminates the decoded signal and decodes the lower-power signal. Hence, the achievable rates of the two users are expressed as follows:

$$R_1 = \log_2(1 + \frac{P_1|h_1|^2}{P_2|h_2|^2+N_0}) \tag{2.5}$$

$$R_2 = \log_2(1 + \frac{P_2|h_2|^2}{N_0}) \tag{2.6}$$

where user 1 is assumed to be the one whose received power is the highest at the BS, with transmission power P_1. Also, P_2 is the transmission power of user 2. Figure 2.6 illustrates the capacity regions for the cases of symmetric and asymmetric channels, respectively. The figure shows that NOMA is capable of achieving the uplink capacity region where the different points are dependent on the users decoding order. In contrast, OMA is shown to be suboptimal in general for both symmetric and asymmetric channels, except for one point. However, in asym-

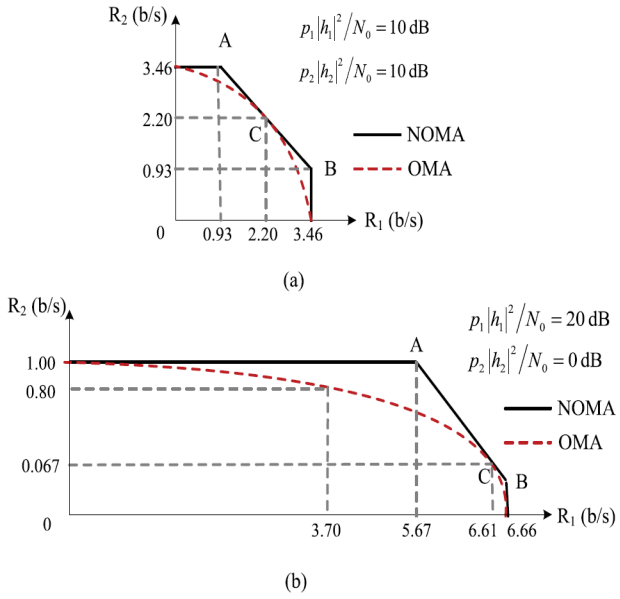

Figure 2.6: Throughput (capacity) region in uplink [65]

metric case, this optimal point suffers from sever fairness problem where most of the bandwidth is allocated to the user with better channel conditions to attain the maximum throughput. The results of both downlink and uplink capacity regions imply that NOMA has the potential to attain better balance between system efficiency and fairness.

2.4.3 Power-Domain NOMA

In power-domain (PD) NOMA, multiple access is realized by allocating different power levels to each NOMA user sharing the same orthogonal resource. The power allocation mainly depends on the channel gains of the users such that distinction can be attained at the receiver. At the receiver side, the power difference is utilized in distinguishing the signals of the superimposed users whereas SIC is commonly used. According to the received signal-to-interference-plus-noise ratio (SINR), SIC is carried out at a descending order. Particularly, the SIC receiver decodes the strongest signal while treating all other weaker signals as interference. Then, the successfully decoded signal is subtracted and the process is repeated until all the superimposed messages are decoded [64].

As mentioned, in the downlink, the power split is crucial for the achievable rates of the users. In order to maximize the network throughput, higher power should be allocated to the users with better channel conditions. However, this increases the risk of outage encountered by the users

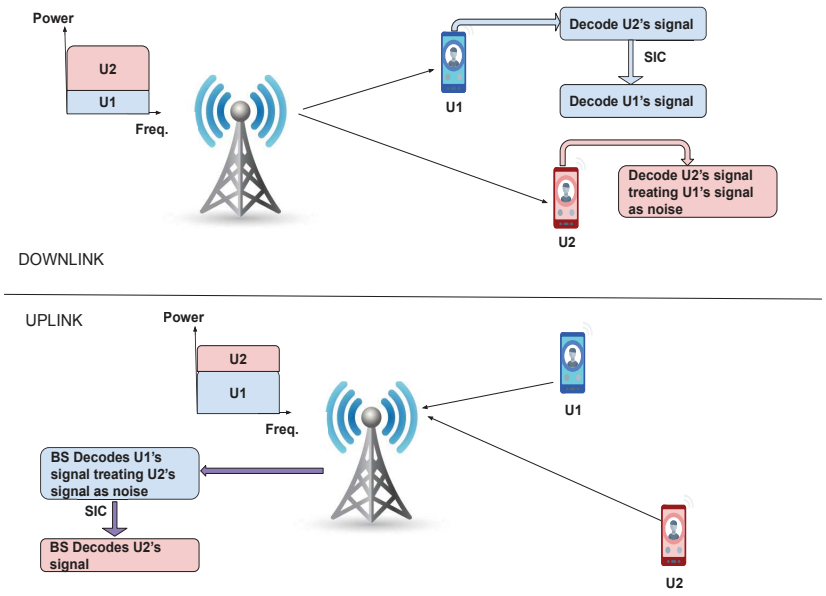

Figure 2.7: Basic 2-user PD-NOMA with SIC in both downlink (upper) and uplink (lower)

having poor channel conditions. To approach fairness, the power allocation should be inversely proportional to the channel gains [11]. In uplink, the transmission powers should be allocated based on the the users' channel conditions keeping in mind that an enhanced performance gain can be attained by clustering more distinctive users. Figure 2.7 clarifies the basic PD-NOMA scheme for 2-users with SIC decoder. For uplink NOMA, we assume that the decoding is done for the user with better channel conditions first. Otherwise, the user with poor channel conditions would need significant amount of transmission power to compensate the severe channel attenuation [66].

PD-NOMA has a number of related challenges. In particular, the clustering of the users (referred to as pairing for 2-user NOMA) is a major challenge. The users clustering refers to the decision on which users should share the same resource. This can be done based on different metrics such as distances, channel conditions, level of sensed interference, and users' QoS requirements [28]. Generally, the optimal NOMA clustering is not an easy-to-solve problem that requires exhaustive search [67]. Particularly, for each user, one needs to find all possible combinations of user grouping. Consequently, it may not be practical specially for systems with large number of users [68]. Therefore, heuristic low-complexity solutions are usually used to solve the problem. Alternatively, game theory [69], matching theory [70, 71], and relaxed optimization approaches were employed to solve the problem [67]. Also, the SIC decoding order of the clustered users is another challenge that affects the NOMA performance [72]. In litera-

ture, several works adopted PD-NOMA schemes. For instance, the 3GPP considered NOMA schemes under the name of multi-user superposition transmission (MUST) in Release 13 [73]. In [74], a multi-cell NOMA-based wireless system was studied considering both uplink and downlink directions. Particularly, the authors developed an analytical framework to asses the outage probability and the achievable rates of the system. In [75], the optimal power allocation as well as the max-min achievable rates of two and three user cooperative NOMA schemes were investigated. It is worth pointing that user-cooperative PD-NOMA systems allows the strong user that decodes the weaker users' signals, to act as a relay to enhance the data rates of the weak users.

An essential variant of PD-NOMA referred to as cognitive radio (CR) NOMA was provided in literature [11]. Unlike the conventional NOMA case, CR-NOMA strictly ensures that the QoS of some or all of the NOMA users are met. In other words, in CR-NOMA, NOMA is considered as special case of cognitive radio where the power allocation is done based on the predefined QoS requirements of the users. In [76], the authors applied CR-NOMA in a 2-user downlink network where the user with poor channel conditions was considered as the primary user that has predefined rate requirements. Thus, the power allocation is done to achieve this requirement where most of the power is gone to this poor user. The remaining power, if any, is allocated to the other user. The rational behind this strategy is to allow other users to share the spectrum with the poor user (i.e primary) that must be served for priority, or fairness issues. This definitely brings considerable gains as compared to OMA in terms of sum rates and spectrum efficiency. In [66], a generalized power allocation framework for both the downlink and uplink of CR-NOMA networks is studied. The objective is to meet all the users QoS requirements in a more flexible manner to realize different tradeoffs between users' fairness and system throughput. A different perspective of CR-NOMA can be found in [77] where a NOMA-based secondary network is assumed to operate with an existing primary network. In this regard, CR-NOMA aims to achieve enhanced utilization of the available spectrum. A comprehensive survey of works applied this CR-NOMA perspective can be found in [67].

2.4.4 Code-Domain NOMA

In code-domain (CD) NOMA, different users are multiplexed at the same frequency-time resource by allocating them different spreading sequences. Unlike CDMA, the spreading sequences of CD-NOMA are restricted to low-density (sparse) sequences or non-orthogonal se-

quences with low cross-correlation [16]. A major difference between PD-NOMA and CD-NOMA is that CD-NOMA can achieve a spreading gain and shaping gain at the cost of more consumed bandwidth [16, 74].

Several techniques are within the frame of CD-NOMA including low-density spreading CDMA (LDS-CDMA), LDS-orthogonal frequency division Multiplexing (LDS-OFDM), sparse code multiple access (SCMA), multi-user shared access (MUSA), and SIC Aided Multiple Access (SAMA), among others [78, 79]. For instance, the key idea behind LDS-CDMA is to employ sparse spreading sequences instead of their dense counterparts used in the conventional CDMA [80]. Thus, with proper design of the spreading sequences, interference can be efficiently mitigated among multiple users. This allows the multiplexing of more users whereas a message passing algorithm (MPA) can be used to realize multiple users detection (MUD) at the receiver side [78]. On the other hand, SCMA is an enhanced version of LDS-CDMA that is capable of achieving an improved average symbol energy known as shaping gain [81]. A recent comparative study between SCMA and dense code multiple access (DCMA) that can be differentiated in terms of their codebooks having sparsity or not, can be found in [79]. In MUSA, each user randomly pick a sequence from an available pool of spreading sequences [82]. Since different spreading sequences may be used for different symbols for the same user, a performance boost can be achieved via interference averaging [78]. At the receiver, codeword-level SIC is used to separate data from different users.

2.4.5 NOMA for MTC

Having the ability to accommodate higher number of users than the available resources, NOMA has a great potential in supporting the massive connectivity requirement of MTC while boosting the spectral efficiency. By allowing multiple MTDs to simultaneously share the same uplink resource either in code or power domain, the access latency can be significantly reduced. However, this comes with the cost of increased receiver complexity to perform inter-user interference cancellation. Fortunately, for uplink NOMA such complex receiver is deployed at the BS which is desirable for low-cost MTDs with mostly-uplink traffic. However, one has to keep in mind, that downlink NOMA with MTC is interpreted as an additional complexity at the MTC devices. This certainly would be challenging such that the MTC devices with limited computational capabilities can perform SIC, for instance.

Efficient NOMA requires the design of advanced pairing and power allocation techniques, as

well as powerful channel coding schemes to ease the decoding task [12,16]. Usually, this is done with the help of the BS which has full information about the users. However, for mMTC gathering such information from massive number of MTDs is impractical. For instance, consider a typical MTC network with N MTDs and M time-frequency resources such that $M << N$. An optimal user pairing algorithm requires the gathering of NM channel coefficients, as well as the QoS requirements of the N users, if any [12]. Such huge amount of information exchange renders the whole system impractical, specifically with grant-free access which is uncoordinated by nature. This poses the challenge of the efficient design of the contention units used by the MTDs for transmissions, in addition to the ability to resolve the collisions among the multiplexed MTDs. In fast-uplink grant, NOMA is promising such that the grant is assigned to multiple MTDs simultaneously. Consequently, the probability of wasting the resources assigned to inactive MTDs could be decreased. However, distributed low-complexity NOMA techniques are also needed to avoid any excessive overhead for information exchange between the MTDs and the BS.

With the high demand of massive connectivity, determining the size of a NOMA cluster is a critical challenge [28]. Specifically, increasing the cluster size allows for higher number of supported devices. However, the efficient decoding of the superimposed signals forces practical constraints on the power level differences in PD-NOMA or the codebook size in CD-NOMA. In addition, the increased cluster size would badly affect the latency requirements, specially for URLLC MTDs. This is because the decoding process will be time consuming to effectively decode large number of superimposed signals. Furthermore, the NOMA receivers complexity increases dramatically with the high increase in the cluster size [78]. Particularly, the SIC-based receiver complexity is $\mathcal{O}(K^3)$, where K is the number of users. This complexity is even much lower than that of the optimal maximum likelihood (ML) detection, whose complexity $\mathcal{O}(|\mathbb{X}|^K)$, where $|\mathbb{X}|$ is the cardinality of the constellation set $\mathbb{X}$. For the MPA-based receiver used in CD-NOMA, the complexity is $\mathcal{O}(|\mathbb{X}|^w)$, where w is the maximum number of non-zero signals superimposed in each chip or subcarrier. Finally, it is worth noting that SIC-based receivers suffers from error propagation, that is more critical when the number of users is sufficiently high. This error propagation means that if the receiver fails to decode certain signal, its interference affects the decoding probability of all remaining signals that should be decoded sequentially. However, some advanced user pairing and power allocation methods, as well as powerful channel coding schemes can be used for reducing the error probability [16]. Keeping

in mind the aforementioned complexity of the NOMA receivers, it is important to control the number of MTDs sharing the same resource to strike balance between system performance and complexity.

2.5 Opportunistic Splitting Algorithm for Wireless Networks

The opportunistic splitting algorithm (OSA) is a distributed medium access control (MAC) protocol originally introduced for resolving collisions of uplink multi-user systems over a sequence of mini-slots [83]. A splitting algorithm is a tree-like mechanism that divides the users involved in a collision into several subsets. Then, only the user(s) in one of the subsets will transmit at the next time slot. Consequently, the probability of collision is reduced. Unlike traditional splitting algorithms, the main objective of OSA is not only resolving collisions, but also finding the user with the best channel gain to be granted access to the contention resource. Therefore, the system's throughput is improved approaching the optimal value as the channel coherence time increases. It shall be emphasized that being "distributed" means that is no need for all users to transmit their channel conditions to a centralized controller (e.g. BS). Instead, each user is assumed to have only knowledge of its own channel conditions, and no knowledge of other users' channels. Hence, the transmission decision is made by each user using its local information.

In our work, we employ a modified version of OSA for resolving collisions among different contending MTDs in an MTC network. In the following, a more in-depth description about the original OSA [83] is given as a background.

OSA Description

The idea behind the OSA is to set two thresholds H_l and H_h and allow only the users whose channel gains are in this range to transmit [83]. It should be noted that both thresholds H_l and H_h can be calculated in a distributed manner. Also, the time slot is divided to a number of mini-slots. Each transmitter receives a feedback packet at the end of each mini-slot. This feedback packet has one of three possible values "0", "1" or "e" corresponding to idle, success or collision. According to the received feedback, the action in the next mini-slot is determined. In case of success, which means that only the best user transmitted in the mini-slot, this winner user transmits until the end of the current time slot. In collision or idle cases the transmitters

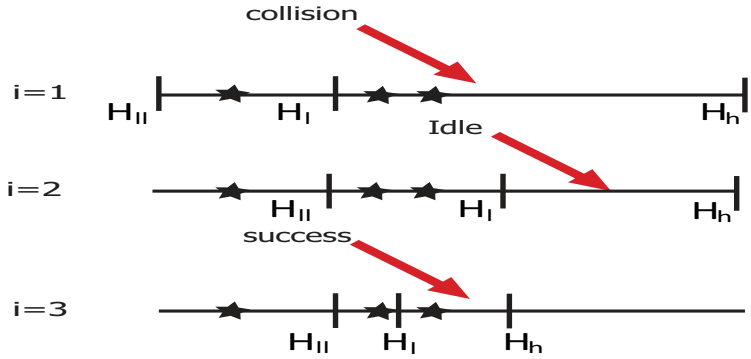

Figure 2.8: Opportunistic splitting algorithm (OSA) procedure in collision case [84]

will adjust the thresholds and continue the algorithm until either a success occurs or the time slot ends. The thresholds updating step is based on two functions which are the lower and split functions. These functions are chosen to minimize the number of mini-slots required to take the decision. Define the complementary cumulative distribution function (CCDF) of the channel gains as $F_H(.)$ and the number of users running the algorithm and competing over the spectrum as η. The two functions are calculated based on $F_H(.)$ and η as:

$$\text{lower}(H_l, \eta) = \begin{cases} F^{-1}(F(H_l)(1-\frac{1}{\eta})+\frac{1}{\eta}) & \text{if } H_l > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{2.7}$$

$$\text{split}(H_l, H_h) = F^{-1}(\frac{F(H_l)+F(H_h)}{2}) \tag{2.8}$$

Algorithm 1 provides pseudo-code that illustrates the procedure followed at each mini-slot until the time slot ends or success occurs. It shall be noted that η is utilized for the initialization of the threshold. Hence, it is assumed that η is known by each user. Figure 2.8 and Figure 2.9 illustrate the OSA procedures in collision case and the case if the first mini-slot is idle respectively [84]. The key aspect of OSA is that the average number of mini-slots required to resolve the collisions and find the user with best channel gain (i.e. take the decision) using OSA is upper bounded by 2.5070 [83]. Generalized versions of OSA that can be applied to a heterogeneous network, where the channel gains of different users are still independent, but may not be identically distributed can be found in [85, 86].

Algorithm 1 Pseudo-code for OSA

Initialize: $H_{ll} = 0$, $H_l = F^{-1}(\frac{1}{\eta})$, $H_h = \infty$
if $H_l \leq |h| < H_h$ **then** Transmit
end if
Receive $feedback \in (0, 1, e)$
if $feedback = e$ **then**
    $H_{ll} = H_l$
    $H_l = \text{split}(H_l, H_h)$
else if $feedback = 0$
    $H_h = H_l$
    if $H_{ll} \neq 0$ **then** ▷ collision before
        $H_l = \text{split}(H_{ll}, H_h)$
    else
        $H_l = \text{lower}(H_l, \eta)$
    end if
end if

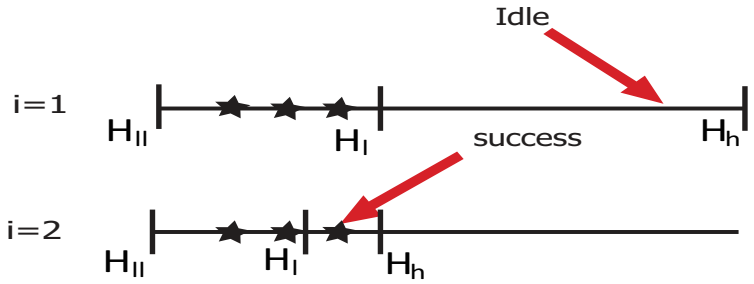

Figure 2.9: Opportunistic splitting algorithm (OSA) procedure if the first mini-slot is idle [84]

2.6 Machine Learning for MTC

In wireless networks, understanding the context of the surrounding environment can effectively facilitate the development of efficient communication protocols while taking optimized decisions [1]. These context information could be related to users activities, QoS requirements, battery levels of the devices, links quality, among others. However, gathering this information could be a challenging tasks in certain communication networks. This is a typical situation of mMTC where gathering all the needed information for performance optimization from massive number of MTDs at the BS is almost impossible. In such environment, machine learning (ML) techniques can prove their superiority. Indeed, ML techniques can be useful in environments where no or limited prior information is available and actions or decisions are needed. In particular, ML are suitable for problems that are too complex to be modeled but have the hidden patterns which can be explored with the assistance of ML [1].

ML techniques were utilized to handle massive access considering both grant-based [1], and grant-free schemes [14]. In [15], the authors indicated a number of potential ML techniques to handle fast uplink grant challenges related to source traffic prediction and optimal scheduling of the MTDs. Applying ML techniques with IoT has a number of associated challenges among which the computation complexity of ML techniques along with the low capabilities of MTDs [1]. Moreover, learning time should be kept as small as possible to be able to support low-latency applications such as eHealth. Based on that ML algorithms should be carefully implemented to comply with MTC unique features and requirements including heterogeneity, and resource constraints. It is worth mentioning that Q-Learning is a simple, model-free reinforcement learning technique that is suitable for low-cost MTDs and hence it was widely used with IoT (e.g. see [1], [87]- [89]). It is also worth pointing that deploying MTDs with artificial intelligent (AI) capabilities such that the devices can learn, infer, and adapt to the environment can effectively boost the performance. However, this comes at the expense of increased cost and computational capabilities of the MTDs.

Reinforcement learning (RL) is a machine learning technique that mimics the fundamental way in which humans learn [90]. Particularly, the learning agent evolves by interacting with the environment through observing it, taking actions and receiving immediate reward feedback. The major goal of the agent is to select actions that maximize cumulative future rewards. RL could also be seen as a trial-and-error technique where an action is decided through learning the system behavior in a given environment [88]. Q-learning is a simplified model of RL that maps situations to actions to maximize a reward. A value function $Q(a, s)$ measures the expected reward from taking certain action, a, being at certain state, s. In literature, different works applied Q-learning techniques to solve the RAN overload problem of mMTC. In [88], a Q-learning RACH (QL-RACH) access scheme was introduced to intelligently assign the access slots to the MTDs sharing the resources with the existing H2H traffic. Furthermore, Q-learning was employed to dynamically adapt the ACB factor in ACB scheme based on the load conditions [91]. To attain further enhancement, the ACB factor adjustment is done with the help of deep RL (DRL) [90]. In this scheme, an offline deep neural network phase is added to the online reinforcement learning phase (i.e.Q-learning) where the devices' QoS requirements were considered. Other proposals utilized Q-learning for preambles division among M2M and H2H traffic [92], or for load balancing via eNB selection [32].

In problems that acquire the best selection policies such as spectrum scheduling and best

relay selections, multi-armed bandit (MAB) learning techniques were usually employed [93]. MAB is essentially an online reinforcement learning technique in which an agent interacts with the environment and learns from its actions [94]. For upcoming 5G networks, efficient scheduling along with full utilization of the spectrum is vital to guarantee the system performance in terms of throughput, delay, and so on. For uncertain and highly competitive environment (such as mMTC), the efficient development of dynamic spectrum access becomes more challenging. In wireless communications, MAB has been adopted for the mode selection, resource allocation, neighbor discovery and selection problems in device-to-device (D2D) communications [95,96]. Additionally, MAB has been utilized for channel assignment in IoT networks [97] as well as the joint channel selection and power allocation in NOMA [98]. For cognitive radio networks (CRNs), opportunistic spectrum access was realized by integrating MAB and matching theory [99, 100]. A comprehensive survey of MAB-based spectrum access mechanisms for wireless networks could be found in [94].

In our proposed work, MAB learning is incorporated to enhance the spectrum access for MTDs, specially in massive scenarios. Hence, in what follows, a brief background about the formulation of MAB problems is provided.

Multi-Armed Bandit (MAB): Problem Formulation and Algorithms

MAB is a class of decision making problems introduced by machine learning the captures the crucial tradeoff between exploration (acquiring new information) and exploitation (optimizing based on the currently-available information) [101]. For instance, in MAB allocation problems, a set of resources is available and it must be allocated to different competing units such that the cumulative system gain is maximized. This allocation is done by a decision maker where the properties of each choice (action) are only partially known at the time of allocation. Then, this knowledge is enhanced as time passes and actions are taken which is basically a reinforcement learning problem.

As indicated in [94, 102], in the most basic formulation, a bandit machine with K arms is considered such that each arm, k, has a reward $X_k(t)$ at time slot t. For simplicity, the rewards could be assumed to be bounded in [0,1]. In the adversarial model one makes no other restrictions on the sequence of rewards. In the stochastic model, the reward $X_k(t)$ for each arm, k, is assumed to be independently and identically drawn from unknown probability distribution, ν_k [101]. Also, the mean reward of arm k is denoted by μ_k (i.e mean of the distribution ν_k),

which is certainly unknown to the decision maker as well.

Generally, at each time slot, an agent plays an arm (action) and observes its associated reward. The performance is usually evaluated in terms of the regret which is defined as the loss for not playing optimally. In other words, the regret is the expected loss of reward due to the fact the played arm may not be the best arm that could have been played due to the lack of information (i.e hidden statistics). The main objective is to design a randomized policy $\Pi = \pi(1), \ldots, \pi(T)$ within time horizon T, where $\pi(t) \in 1, 2, \ldots, K$ denotes the played arm at time slot t, such that the regret is minimized. Employing the policy Π up to time T, the expected regret is defined by [103]:

$$\mathbb{E}R_{\Pi,T} = \mathbb{E}\left[\max_{i=1,\ldots,K} \sum_{t=1}^{T} X_i(t) - \sum_{t=1}^{T} X_{\pi(t)}(t)\right], \tag{2.9}$$

where $\mathbb{E}$ denotes the expectation function that is taken with respect to the random draw of both the rewards and the agent's actions. Another notation for regret is referred to as the pseudo-regret (or weak regret) is defined by [103]:

$$\bar{R}_{\Pi,T} = \max_{i=1,\ldots,K} \mathbb{E}\left[\sum_{t=1}^{T} X_i(t) - \sum_{t=1}^{T} X_{\pi(t)}(t)\right], \tag{2.10}$$

where this formulation is a weaker notation of regret such that the achieved performance is compared to the optimal one in expectation. More formally, one has $\bar{R}_{\Pi,T} \leq \mathbb{E}R_{\Pi,T}$.

To attain the regret minimization goal, the algorithms for MAB problems aim to capture the crucial tradeoff between exploration and exploitation [94], [101]. In one hand, exploration encourages the acquiring of new information by trying out each arm to gain more knowledge about its reward distribution for long-term benefit. On the other hand, exploitation recommends optimizing the actions based on the current available information. Consequently, exploitation means playing the "best" arm according to the current empirical estimates of the rewards to maximize the immediate return. In this regard, one of the simplest and commonly-used strategies to solve MAB problems is the ε-greedy algorithm [102]. Specifically, this algorithm employs an exploration parameter ε such that the decision maker selects a random arm with probability ε. Otherwise, the arm that yields the highest expected reward is selected. Consequently, increasing ε means increasing the exploration done by testing arbitrary arms that could lead to higher regret and slow convergence. Additionally, since ε is constant, the agents continue to choose

random arms even after convergence which leads to a linear increase in the regret function with time [94]. In [98], an enhanced version of this algorithm, referred to as ε-decreasing has been utilized, in which ε decreases with time. This allows the increase of the exploration parameter at the beginning to explore the arms space. Then, the exploration parameter continuously shrinks with time resulting in enhanced regret.

A major drawback of the aforementioned ε-based algorithms is that the exploration of the arms is done in a purely randomized manner. Specifically, the algorithms do not take into account the confidence intervals on the empirical estimates of the arms' rewards [94]. This knowledge is exploited by the upper confidence bound (UCB) algorithms [104]. In the seminal UCB algorithm, an upper-bound for the expected reward is evaluated for each arm. This bound depends on the previously gained rewards by selecting each arm, the total times it has been selected till now, and some uncertainty factor. At each time slot, the UCB algorithm selects the arm, $\pi(t)$, that has the largest UCB index as follows [93]:

$$\pi(t) = \underset{k=1,\dots,K}{\arg\max}\, \bar{X}_k(t) + c_k(t) \tag{2.11}$$

$$\bar{X}_k(t) = \frac{1}{N_k(t)} \sum_{s=1}^{t} X_k(s) \mathbb{1}[\pi(s) = k] \tag{2.12}$$

where $\bar{X}_k(t)$ is the empirical mean of the previous rewards of arm k up to time t. Also, $N_k(t)$ is the number of times the k^{th} arm has been selected. $\mathbb{1}[.]$ is an indicator function that returns 1 is its argument holds and 0 otherwise. The term $c_k(t)$ refers to the padding function that represents the uncertainty of the calculated arms' rewards. Hence, the padding function has higher values for less selected arms. This padding function achieves the required tradeoff between exploration and exploitation. This is done by allowing more exploration of arms with high UCB index due to having high padding value (i.e. more uncertainty). In contrast, the more the arm is played, the smaller its UCB index becomes due to the smaller values of the padding function. This indicates that the empirical mean is closer to the real expected value of the arm. A standard choice for the padding function is [104]:

$$c_k(t) = B\sqrt{\frac{\phi \log t}{N_k(t)}} \tag{2.13}$$

where B is an upper-bound on the rewards and is selected as the maximum value of the observed reward through many trials of the algorithm. ϕ is a parameter that provides a tradeoff between

exploration and exploitation. Particularly, larger values of ϕ result into a higher amount of exploration. It shall be noted also that this choice for the padding function depends on the number of times the arm was played $N_k(t)$. This dependency also serves the exploration and exploitation tradeoff by lowering the UCB values of the more selected arms. Using the UCB algorithm, the expected regret can be upper-bounded by $O(logT)$ [94]. For the adversarial MAB problems, another family of approaches referred to as exponential-weight algorithms can also be used to solve the problem [102]- [103].

2.7 Summary

In this chapter, we gave the necessary background information related to the topics studied in this thesis. First, an overview of the IoT along with M2M communications was given. In this regard, we provided an overview of supporting technologies, as well as the main features and challenges of M2M communications. We also shed light on some of the standardization efforts made by the 3GPP for cellular MTC adaptation. Next, we provided an overview of different uplink access techniques employed for MTC. This was also supported by a discussion of the associated challenges and some of the state-of-the-art proposals for each access technique. Moreover, NOMA technique was briefly surveyed showing its potential for supporting MTC specially with massive access. Afterwards, we reviewed the OSA algorithm that is utilized in our proposed work of chapters 3 and 4 for providing distributed access for massive MTC. We also shed light on the potential of powerful ML tools to solve the problems inherited in the different reviewed access techniques for mMTC. In chapters 5 and 6, we propose an uplink access technique for MTC that employs the fast uplink grant scheme. In this regard, MAB learning is employed to efficiently tackle the scheduling challenge, and hence a thoroughly review of MAB learning was provided in this chapter.

3. A Scalable Overload Control Algorithm for Massive Access in M2M Networks

3.1 Introduction

With the emergence of Internet of Things (IoT) and its applications that are going to formulate a paradigm shift in the wireless communications area in the near future, machine-to-machine (M2M) communication has recently received special attention as the main platform for IoT [37]. M2M communication, also referred to as machine-type communication (MTC), provides automated data transmissions among machines with little or no human intervention. M2M communication has a great potential to be utilized in different fields and support different applications such as eHealth, intelligent transportation systems, smart cities, and smart grids, among others.

The large-scale deployment of MTC devices is the main source of radio access network (RAN) congestion and overload. Particularly, congestion occurs when enormous number of MTC devices access the network simultaneously or even in small period of time. For instance, this situation is likely to happen after power failure. As a result, the devices could suffer from severe access delay, packet loss, and resource wastage which are all undesired for energy-constrained MTC devices. Hence, efficient overload control mechanisms are essential for successful operation of M2M networks. In literature, random access mechanisms are commonly adopted for MTC access and connection establishment. The reason is that random access mechanisms are simple, flexible and could be done by the devices independently without central coordinator [29]. Even though cellular networks are promising for supporting MTC, the traditional random access (RA) protocols used in LTE-A are inefficient for MTC [105]. With massive access, these protocols suffer from increased collisions that seriously degrade the network performance. Several modifications were applied to improve the traditional LTE-A random access performance. For instance, the 3GPP has identified a number of candidate overload

solutions such as access class barring (ACB) scheme, backoff scheme, and pull-based scheme, among others [6]. However, these proposals achieve limited enhancements and collisions remain the performance bottleneck for MTC.

In this chapter, we propose a network overload solution that is capable of overcoming the increased collisions resulted from the RA schemes used in LTE-A. The proposed algorithm, referred to as M2M-OSA, is an extension of the opportunistic splitting algorithm (OSA) proposed in [83], [86], adapted to M2M communications. The major idea of M2M-OSA is to allow the active devices to draw random numbers and enter a competition for the available resources if this number is between two prespecified thresholds. The M2M-OSA can handle large number of users and allocate the resources among them with low overhead and delay in a distributed manner. That is the major requirements of MTC networks (e.g. scalability, low overhead and delay) are all satisfied with low complexity. Additionally, instead of just sending a request, the algorithm allows for fast data transmission right after gaining access to the network which is more suitable for the MTC devices with low data transmissions. Moreover, we devise a method for the estimation of the number of active MTDs at each time slot. Finally, we discuss a number of non-ideal situations that could be encountered in real-life deployment of the M2M network.

3.2 Related Works

In this section, we review several proposals that either applied modifications to or avoided the 4-way conventional RA while handling M2M massive access.

In LTE-A networks, two types of RA are known: contention-based and contention-free. In the following, we focus on the contention-based RA procedure which is usually used with MTC due to the enormous number of devices in the network. The UEs are allowed to perform RA during specified periods called RA slots. Also, the resources dedicated by the eNB for the RA process are divided between two channels, namely physical random access channel (PRACH) and physical uplink shared channel (PUSCH). The contention-based RA procedure is a 4-step handshaking starts by RA preamble transmission and is finalized by a contention resolution message as detailed in Section 2.3.1.1. During the first step, multiple UEs can select the same preamble and transmit it. Hence, collision occurs and the resource is wasted.

In overload situations, numerous number of MTC devices attempt to access the network concurrently. This results in network congestion that leads to severe delay, packet loss and even

service interruption. Efficient overload control mechanisms are needed to ensure the service availability and meet the performance requirements. A number of candidate solutions have been identified by the 3GPP such as ACB, backoff scheme, dynamic allocation of PRACH resources, separate PRACH resources for MTC, and pull-based scheme [6]. The idea behind such schemes is to control the number of contending devices and reduce the interaction among them. For instance, the backoff scheme delays the random access reattempts of the failed devices. However, it is only effective for light load conditions. In ACB schemes [106]- [108], the eNB announces access probabilities (AP), known as ACB factor, to the devices. Then, each device draws a random number and only starts the RA if this number is less than the AP. Extended-access barring (EAB) scheme was also introduced to restrict access of delay-tolerant devices. The 3GPP recommended the use of ACB schemes, and particularly defined EAB as feasible solution for overload. However, ACB schemes could result in longer delays to some devices.

Several proposals applied different modifications to enhance the RA performance. In [107], the ACB scheme was improved by dynamically adapting the ACB factor. In dynamic ACB (DACB), the ACB factor is adjusted based on the load conditions. It was shown that the optimal performance depends on the number of backlogged UEs. Hence, a heuristic algorithm was proposed to estimate the load and approach the optimal performance. In [108], a two-stage ACB check was proposed to allocate the uplink resources where in the first stage DACB is performed. Then, the failed MTC devices in the first stage can perform a second ACB check and be allocated the remaining uplink resources. In [109], an analytical framework for the RA scheme was proposed and the access performance was optimized by tuning the ACB factor and the uniform back-off window size parameters. In [58], the collisions resulted from RA are reduced by allocating the same preamble set to groups of devices that are far apart. Hence, depending on the propagation delays the eNB can detect identical preamble by multiple users and send different RAR messages. The scheme results in less delay but the collisions within each group remain. More proposals that applied modifications to the 4-step RA could be found in [30, 110, 111].

Although utilizing the 4-step RA is popular in handling massive MTC access, it was avoided in some proposals due to its vulnerability to collisions [9, 112, 113]. For instance, a hybrid medium-access control protocol that employs carrier-sense multiple-access (CSMA) technique for contention and then allocate the resources to the winner devices in a time-division multiple-access (TDMA) fashion was introduced in [112]. This scheme introduces a trade-off between

the performance of both CSMA and TDMA, and requires a carrier sensing module at the MTC devices. In [113], an adaptive multi-channel protocol that divides a frame into negotiation and transmission periods was proposed. The protocol divides a frame into negotiation and transmission periods, and optimizes the negotiation period and access probability such that the channel utilization is maximized. The negotiations are done via a common control channel (CCC). Hence, the successes are decided serially which increases the delay and wastage of the data-channels resource. Using code-division multiple-access (CDMA), a distributed queue (DQ) approach is introduced in [9] to eliminate the RA attempts. The DQ approach depends on elected devices (leaders) to receive initial requests from other devices and organize them in a virtual queue. The devices then access the network based on their position on the queue. However, this approach requires more complicated hardware and excess computational effort by assuming leaders with multiple packet reception (MPR) and successive interference cancellation (SIC) capabilities. In the following, we describe the proposed M2M-OSA that avoids all the problems indicated above.

3.3 System Model

Consider the arrival of N MTC devices (MTDs) to a single cell within a bounded activation interval $[0, T_A]$ to be served by a single BS. Assuming that the devices access the network in a highly synchronized manner, the Beta probability distribution is used to model the probability of the devices activation with parameters $(\alpha = 3, \beta = 4)$ [6]. We assume that a device is active when it has data for transmission. Also let the activation time T_A be divided into I_A time slots. Based on that, each device is activated at time $t \in [0, T_A]$ with probability $g(t)$ calculated as:

$$g(t) = \frac{t^{\alpha-1}(T_A - t)^{\beta-1}}{T_A^{\alpha+\beta-1}\mathcal{B}(\alpha, \beta)}, \tag{3.1}$$

where $\mathcal{B}(\alpha, \beta)$ is the Beta function [114]. Also, we assume that all newly activated devices within a time slot join the backlogged pool and will only attempt accessing the channel at the beginning of the next time slot.

Regarding the frequency resources, we consider the availability of M equally-sized uplink resources for M2M use. Each uplink resource could consist of one/multiple resource blocks (RBs) depending on the M2M application. However, for referral simplicity, we would refer to each uplink resource as RB. We also assume that both H2H and M2M have separate resources,

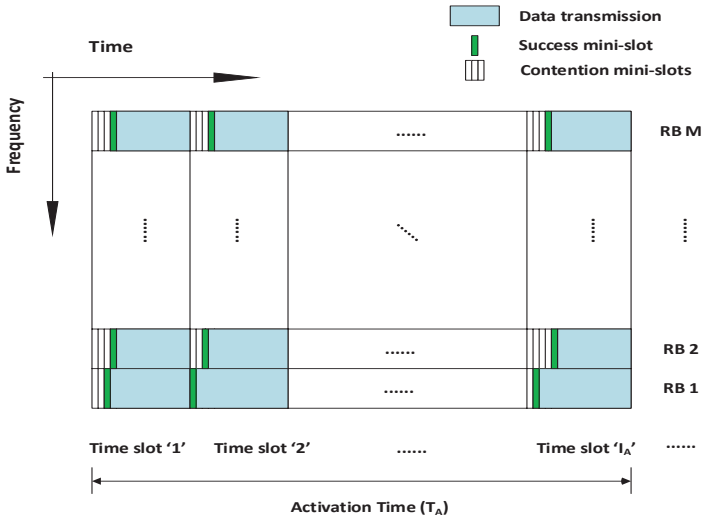

Figure 3.1: The structure of the time/frequency resources of the considered M2M network

and thus, no collision between them. According to 3GPP, usually MTC devices have small amount of data [3]. Hence, we adopt fast data transmission right after successfully granted access [56]. It is worth mentioning that, unlike the conventional RA, in our model all the M resources are utilized for uplink transmission and none of them are wasted in the PRACH.

For the time resources, we consider that each time slot is divided into ζ mini-slots where a mini-slot is the basic time unit for the proposed algorithm. The mini-slot time is assumed to be enough to transmit a small packet and receive its feedback including a guard period to account for different propagation delays. Generally, each time slot consists of two different periods: reservation period and data transmission period. In the reservation period, the active devices contend to get access to the network and transmit their data. In the transmission period, the winner device transmits its data on a preselected uplink resource for the remaining of the time slot. Hence, it is important to keep the number of reservation mini-slots as small as possible to allow for more mini-slots for data transmission. Figure 3.1 illustrates the adopted structure of the time/frequency resources in our proposed algorithm. Focusing on the effect of the proposed access algorithm in providing successful access to MTC devices, ideal channel conditions are assumed [34]. In particular, once a winner is selected, successful access is considered and the physical layer impairments are neglected [53].

3.4 M2M-OSA Procedure

In the proposed algorithm, each active device randomly selects one of the available uplink resources (i.e. RBs) to compete for and get a chance to transmit its data. Then, the devices selecting the same RB apply opportunistic splitting to reach a winner device in a distributed manner. This winner device transmits its data until the current time slot ends, whereas other failed devices shall retry in the next time slot, if permitted. Particularly, the BS broadcasts a maximum retransmission limit, O_m, allowed for each device, if exceeded, the device is blocked. Hence, each device keeps tracking a counter, referred to as overload indicator O, that counts the number of attempts made by the device. The value of O is incremented by one with each failure. At the beginning of each time slot, both the newly activated devices and the backlogged ones with $O \leq O_m$ start to contend for the M RBs. This procedure is repeated until all N devices either successfully access the network or declared blocked.

During the reservation period, each active device draws a random number from the standard uniform distribution (i.e. $U[0, 1]$), and use this number as its competing metric, c. Defining $F(.)$ as the complementary cumulative distribution function (CCDF), each device n should calculate the value of $X_n = F_C(c_n) = p(C > c_n)$ corresponding to its competing metric, c_n. Knowing the values of two thresholds X_l and X_h at certain RB, the devices check if their X values satisfy the condition that $X_l < X < X_h$. If the condition is satisfied, they transmit small request packets (transmission requests) to the BS. The BS replies with a feedback message either $0, 1$ or e, corresponding to idle, success, or collision, respectively. Based on the received feedback, each device independently adjusts the values of the thresholds. The process is repeated till reaching a success or the time slot ends.

Regarding the thresholds updating step, let us define η as the number of devices contending for the same RB. Using η, two functions, namely, upper and split functions are used to adjust the values of the thresholds during contention [86]. These functions are selected such that the probability of success at each mini-slot is maximized [83], and are calculated as follows:

$$\text{upper}(X_h, \eta) = \begin{cases} X_h(1 - \frac{1}{\eta}) + \frac{1}{\eta} & \text{if } X_h < 1, \\ 1 & \text{otherwise} \end{cases} \tag{3.2}$$

$$\text{split}(X_l, X_h) = (\frac{X_l + X_h}{2}) \tag{3.3}$$

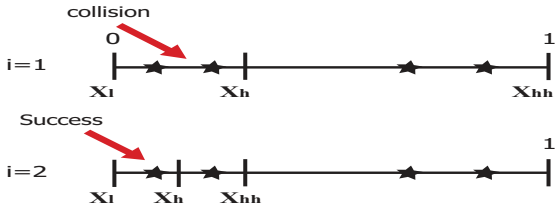

Figure 3.2: Splitting algorithm procedure in collision case

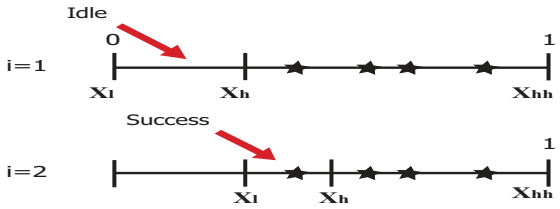

Figure 3.3: Splitting algorithm procedure in case the first mini-slot is idle

An additional threshold, X_{hh}, is used to upper bound the splitting range by 1 as the different values of X correspond to probabilities. Figure 3.2, and Figure 3.3 illustrate the splitting algorithm procedure in collision case, and in case the first mini-slot is idle where upper function is employed, respectively. In addition, Algorithm 2 describes the pseudo-code of the proposed M2M-OSA performed by each active device, n, in a distributed manner at the beginning of each time slot.

The procedure in Algorithm 2 selects the device with the minimum X. Since the selection criterion based on homogeneous random competing metrics all drawn from same distribution independent of the devices parameters (e.g. location), no fairness problems among the devices could be encountered. Also, it was proven that the winner selection procedure using OSA is capable of reaching a decision (i.e. selecting the winner device) in relatively small portion of time provided that both η and $F_C(.)$ are known. In particular, the average number of mini-slots required for the decision τ is upper bounded by 2.5070 regardless the number of contending devices [83]. This result is essentially important for M2M communications where massive access with reduced delay and collisions which in turn results in reduced energy consumption and overhead are crucial requirements.

3.5 Active Devices per Time Slot Estimation

As shown in the previous subsection, the number of devices contending for each RB, η, plays an essential role in the process of adjusting the thresholds of the M2M-OSA. However, in reality

Algorithm 2 Pseudo-code for M2M-OSA Done by Device n

Given: ζ, η
if $O_n \leq O_m$ **then**
  Randomly select a RB to compete for
  Draw a random number $c_n \sim U[0,1]$
  $X_n = F_C(c_n)$
  Initialize: $i = 1$, $X_l = 0$, $X_h = \frac{1}{\eta}$, $X_{hh} = 1$
  while ($i \leq \zeta$ and $feedback \neq 1$) **do**
    if $X_l < X_n < X_h$ **then**
      Transmit the request
    end if
    Receive $feedback \in \{0, 1, e\}$
    if $feedback = e$ **then**
      $X_{hh} = X_h$
      $X_h = \text{split}(X_l, X_h)$
    else if $feedback = 0$ **then**
      $X_l = X_h$
      if $X_{hh} \neq 1$ **then** ▷ collision before
        $X_h = \text{split}(X_l, X_{hh})$
      else
        $X_h = \text{upper}(X_h, \eta)$
      end if
    end if
    $i = i + 1$
  end while
  if $feedback = 1$ **then**
    Data transmission
  else
    $O_n = O_n + 1$
  end if
else
  Blocked
end if

the exact number of active devices at each time slot is unknown to the system (both BS and devices). Therefore, it is important to devise a method to estimate this number. In particular, at each time slot, the number of contending devices is the sum of the new arrivals and the backlogged users from previous slots. Also, assuming no maximum retransmission limit, the number of backlogged devices is the difference between the number of active devices and the number of successes in the previous time slot. Thus, let N_k be the number of active devices at the k^{th} time slot:

$$N_k = \begin{cases} N_{k-1} - S_{k-1} + a_k & \text{if } k \leq I_A, \\ N_{k-1} - S_{k-1} & \text{otherwise,} \end{cases} \tag{3.4}$$

where S_k and a_k are the number of successes and new arrivals at the k^{th} time slot, respectively. Knowing N [6], the expected value of N_k is a reasonable estimate of the actual value. First, the expected number of new arrivals at the k^{th} time slot, λ_k, can be calculated as follows:

$$\lambda_k = \bar{a}_k = N \int_{t_{k-1}}^{t_k} g(t)\mathrm{d}t, \qquad k = 1, 2, \ldots, I_A, \tag{3.5}$$

where t_{k-1} and t_k mark the start and the end of the k^{th} time slot, respectively. Also, $\bar{(.)}$ indicates the expected value. Using the trapezoidal rule [115], the integral in (3.5) can be approximated by:

$$\lambda_k = N(t_k - t_{k-1})\frac{g(t_{k-1}) + g(t_k)}{2}, \qquad k = 1, 2, \ldots, I_A. \tag{3.6}$$

For the average number of successes, $\bar{S}_k$, using M2M-OSA, a success occurs when at least one device contends for a RB[1]. The reason is that the splitting algorithm restricts the average reservation period to at most 2.507 mini-slots regardless of the number of contending devices [83]. This result is true provided that both η and $F_C(.)$ are known, or even an estimate for the average η not the actual value is known as was shown in [84]. Since each device randomly selects one contention RB, the N_k devices are uniformly distributed among the M RBs. Hence, by defining $N_{k,m}$ as the number of devices at the m^{th} RB at the k^{th} time slot, the probability of

[1]This is valid under the assumption that ζ is at reasonable value to accommodate both reservation and transmission.

success at the m^{th} RB, P_{s_m}, and hence $\bar{S}_k$ can be obtained as:

$$N_{k,m} \sim \text{Bin}(N_k, 1/M) \tag{3.7}$$

$$P_{s_m} = p(N_{k,m} \geq 1) = 1 - \left(1 - \frac{1}{M}\right)^{N_k}$$

$$\bar{S}_k = MP_{s_m} = M\left(1 - \left(1 - \frac{1}{M}\right)^{N_k}\right). \tag{3.8}$$

We assume that the BS performs the calculation of the estimated number of devices, $\bar{N}_k$, and broadcasts this information at the beginning of each time slot as follows:

$$\bar{N}_k = \begin{cases} \bar{N}_{k-1} - \bar{S}_{k-1} + \lambda_k & \text{if } k \leq I_A, \\ \bar{N}_{k-1} - \bar{S}_{k-1} & \text{otherwise.} \end{cases} \tag{3.9}$$

Using (3.7) and (3.9), the average number of devices per RB at the k^{th} time slot, η_k, is given by:

$$\eta_k = \frac{\bar{N}_k}{M}. \tag{3.10}$$

Each device uses η_k to adjust the thresholds in the M2M-OSA described in Algorithm 2.

3.6 Effect of System Non-idealities

In this section, we discuss the effect of some non-idealities on the performance of M2M-OSA.

3.6.1 Effect of Errors

Two types of errors are discussed (i) misestimation of the number of devices per time slot and (ii) feedback errors.

3.6.1.1 Misestimation of the number of devices per time slot

We assume that such an error could occur due to inaccurate information about N. This affects the thresholds updating procedure of M2M-OSA. To that end, we discuss two different

situations that result from the following definition of error:

$$\begin{aligned} \text{Err} &= \frac{|N - N_{err}|}{N} \\ N_{err} &= N(1 \mp \text{Err}) \end{aligned} \tag{3.11}$$

where N_{err} is the erroneous value of N known at the BS. The $\mp$ in (3.11) refers to the under-estimation and over-estimation error cases, respectively. For the over-estimation error, the BS assumes $N_{err} > N$ which allows the continuation of the algorithm until serving the actual number of devices. In contrast, the under-estimation case ($N_{err} < N$) leads to a drop in $\bar{N}_k$ faster than the actual N_k. Hence, at the end when $\bar{N}_k$ approaches zero, the algorithm does not work properly where no more devices are served and all the time slots are lost. Particularly, the BS assumes that all the devices were served, whereas in fact there are more devices need service. From this perspective, it may be better for the eNB to over-estimate N. However, this situation could be solved if one allows the BS to maintain the last value of $\bar{N}_k > 0$ until all the devices are served.

In both cases, since the step size of the algorithm is proportional to $1/\eta$, with the mis-estimation of η, there is no guarantee for the 2.507 upper bound on the average reservation period. For the over-estimation case, smaller step size is used as compared to the correct step. Hence the number of collisions is less than the idle mini-slots. The opposite is noticed for the under-estimation case with larger step size. This also will have impact on the average energy consumption of the devices.

3.6.1.2 Feedback error

We focus on two situations where either an idle or success events are mistaken as collisions [116]. For an idle mistaken by collision, the algorithm continues splitting in a range where no devices exist and no success is achieved in this time slot. This situation is known as deadlock. However, usually this problem could be solved by limiting the number of consecutive idle slots and then reinitializing the algorithm [117]. Of course, this would affect the throughput of the algorithm where the upper-bound for the average reservation period could be violated. However, in the context of MTC, the devices usually have small amount of data with strict requirements on key metrics such as access delay and energy but they can tolerate the requirement of high data rate. Hence, monitoring the idle slots is acceptable and effective technique in the studied

Table 3.1: System model parameters

Parameter	Definition	Value
M	Number of uplink resources	15
N	Number of MTC devices	1000:1000:30000
I_A	Number of slots per activation period	100
O_m	Maximum retransmissions limit	100, 300, ∞
ζ	Number of mini-slots per time slot	40

case. For a success mistaken by collision, no modification is needed. However, larger number of reservation mini-slots would be consumed to overcome this error.

3.6.2 Short Time Slot

Up to this point, we have assumed that once a success occurs the remaining transmission time is enough to accommodate the data transmission. This assumption relies on the relatively small bounded average reservation time achieved by the proposed algorithm, and the small-sized packets transmitted by M2M applications. However, in practice, this situation could be violated. To account for this case, we propose a modification that could be applied to M2M-OSA. Particularly, we allow a successful device to continue its data transmission using the same RB for the next time slot, if needed, without the need for another reservation. In such a case, the BS declares this RB as "BUSY" (i.e. unavailable for contention). Upon completion of data transmission, the BS sends a "RELEASE" signal to announce that the busy RB is available. One drawback of this method is that the remaining of the extended time slot(s) on which the device resumes transmission may be lost since the device may not need the whole period. However, it is better to work on time-slot basis to ease the synchronization among all devices. Another suggested approach when the data packets are large is to assume that each device contends for a set of RBs or that a sub-channel consists of multiple RBs (e.g. x) that would be enough to accommodate the full packet transmission. In this case, the number of resources available for M2M use would be M/x sub-channels instead of M. However, this is not the one adopted in this proposal.

3.7 Simulations and Discussions

3.7.1 Simulations Settings and Evaluation Metrics

In this section, we consider a single cell scenario such that the number of devices varies from 1000 to 30000 MTDs. A Beta-distributed traffic model is simulated with ($\alpha = 3, \beta = 4$), and we consider a device is active if it has data to be transmitted. The performance of the proposed algorithm (M2M-OSA) is evaluated compared to the dynamic access class baring (DACB) scheme with optimal ACB factor AP [8]. In DACB, the BS dynamically broadcasts the optimal $AP = \min(1, M/N_k)$ and only the MTDs whose random draws are less than AP will perform the 4-step RA procedure. The DACB scheme assumes zero backoff interval such that the failed devices will reattempt in the next time slot. Also, for fair comparison, it is assumed that no preamble loss due to channel impairments and that a successful preamble always results in successful access. Table 3.1 shows the simulation parameters. We examine the results in both cases of O_m, infinite and finite.

In addition, four key performance metrics were studied to show the effectiveness of the proposed algorithm:

- **Decision delay:** this is the part of each time slot wasted to decide the winner MTD at different RBs. τ is a measure for the average of this time in terms of the number of mini-slots. Reducing the decision delay means more time for data transmission and higher throughput.

- **Total service time (TST):** total time (or number of time slots) needed for all MTC devices to reach a final state either by successfully accessing the channel or declared blocked.

- **Average access delay:** this is the time elapsed from the first access attempt of a device until successful access. In the following, we measure it in terms of the number of consumed time slots by each device before gaining access. Additionally, only successful attempts are considered in calculating the average (i.e. blocked MTDs were ignored).

- **Blocking probability:** it is the ratio between the number of the MTDs that failed to get channel access due to reaching the maximum limit of retransmissions, O_m, to the number of MTDs in the system attempting to get access N.

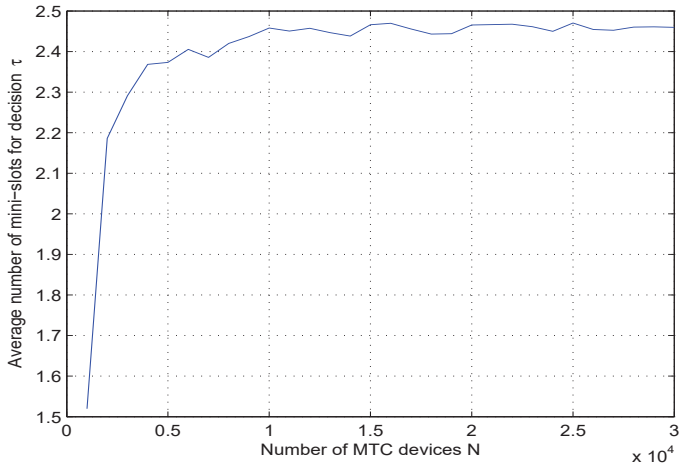

Figure 3.4: Average number of mini-slots required for a decision using M2M-OSA

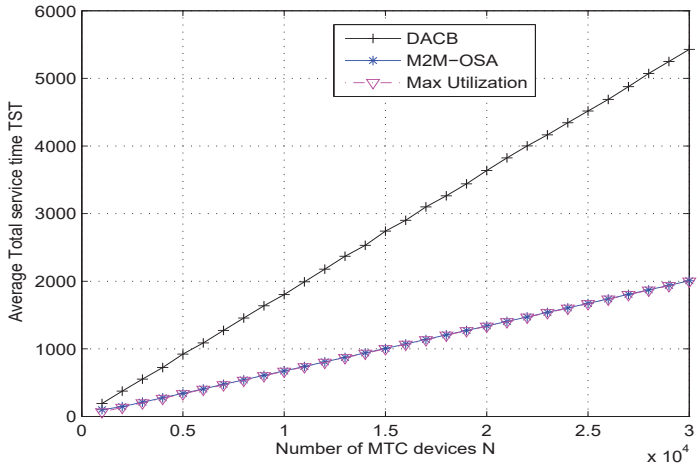

Figure 3.5: Total service time vs. number of MTC devices with no maximum retransmission limit

3.7.2 Simulation Results of M2M-OSA

In this section, we consider the ideal situation such that the number of active MTDs per cycle N_k, and hence η_k are perfectly known. Relaxing this assumption is going to be investigated in upcoming sections. Assuming unlimited persistent scheme which means that no maximum retransmissions limit and that the devices keep trying until gaining successful access (i.e. $O_m \rightarrow \infty$), Figure 3.4 depicts that τ is upper bounded by 2.5 regardless of the massive increase in the number of MTC devices. That is the proposed algorithm does not introduce much delay for the winner selection. Additionally, the proposed algorithm does not result in high loss compared to centralized schedulers that allocate resources to the MTDs without contention but with increased overhead.

Figure 3.5 plots the total service time (TST) in terms of the number of time slots versus the number of MTC devices N in the system. As shown from these results, the proposed distributed technique results in much lower TST in comparison with the optimal DACB scheme. The improved performance of the proposed M2M-OSA is due to the fact that it does not suffer from the increased collisions resulted from the random access mechanism used in DACB. Instead, it guarantees that finally each RB will be allocated to a certain MTD in a time bounded in average by 2.5 mini-slots. That is, if enough MTDs are active, which is the case for M2M with massive access, no RB is wasted. This can be further illustrated by comparing the resulted TST from M2M-OSA with the maximum utilization curve shown in Figure 3.5. The maximum utilization curve is drawn by dividing the number of MTC devices N by the number of available resources M which represents the case of no wasted resources at all. As seen from the results, the proposed scheme offers almost the same performance as the case of maximum utilization. This confirms that M2M-OSA is able to achieve full resource utilization in a distributed manner with low overhead and decision delay. Thus, the proposed algorithm supports network scalability.

Regarding the access delay, Figure 3.6 shows the average access delay resulted from both M2M-OSA and optimal DACB with different values of N and infinite O_m (worst case of collisions). The figure depicts the superiority of M2M-OSA due to its capability to solve the collisions problem that is the main source of congestion in the conventional random access procedure and also in DACB. It can be noticed that the performance gap between the two techniques increases with the increase of N. Consequently, the proposed algorithm provides better adaptation to network scalability, whereas DACB suffers from increased collisions and resource wastage. It is worthy to mention that decreasing the average access delay has a direct impact on reducing the energy consumption which is a key requirement for battery-constrained M2M communications.

If a finite maximum retransmissions limit O_m is enforced by the network, some MTDs will be blocked. Assuring low blocking probability is a desired requirement for M2M communications. Figure 3.7 illustrates that M2M-OSA results in enhanced blocking probability in comparison with DACB with $O_m = 100, 300$. Also, it can be seen from the figure that increasing the number of MTC devices N results in an increase in the blocking probability as a result of the increased contention. Additionally, the results in Figure 3.7 depict the effect of increasing O_m on each technique. In particular, increasing O_m results in a decrease in the blocking probability for both techniques. Although more MTDs get involved in the competition, their chance

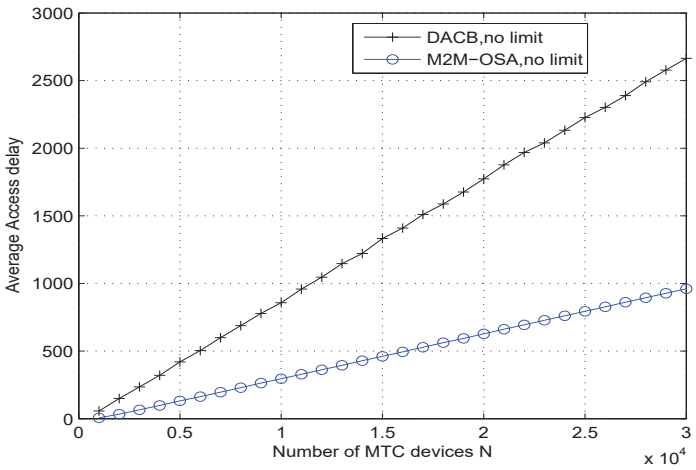

Figure 3.6: Average access delay vs. number of MTC devices for both M2M-OSA and DACB with no maximum retransmission limit

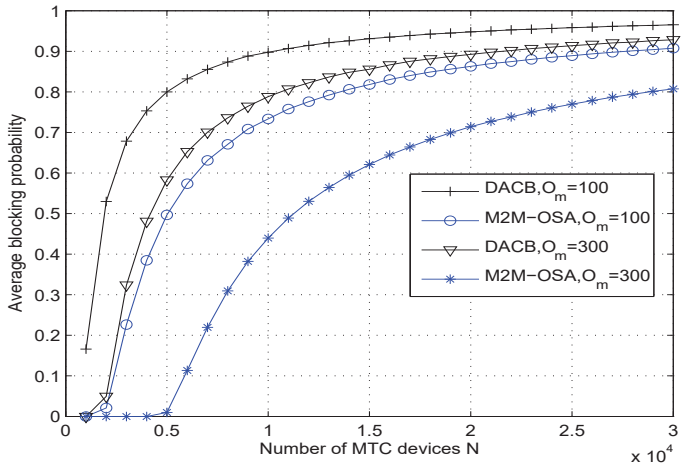

Figure 3.7: Average blocking probability vs. number of MTC devices for both M2M-OSA and DACB with different values of maximum retransmission limit O_m

of getting success increases by allowing more attempts which leads to an improvement in the overall blocking probability of the system. However, this scenario will obviously increase both TST and the average access delay.

To further clarify the effect of the parameter O_m, Figure 3.8 plots the blocking probability vs. $O_m = 50 : 500$ for a network of $N = 8000$. In particular, the figure illustrates the improved blocking probability results from increasing O_m. Furthermore, it shows, along with the result in Figure 3.9, that M2M-OSA is capable of serving more MTDs while consuming less average access delay in comparison to DACB. By examining the curves shown in Figure 3.9 for finite O_m values, it can be noticed that the gap in the performance between the two algorithms decreases

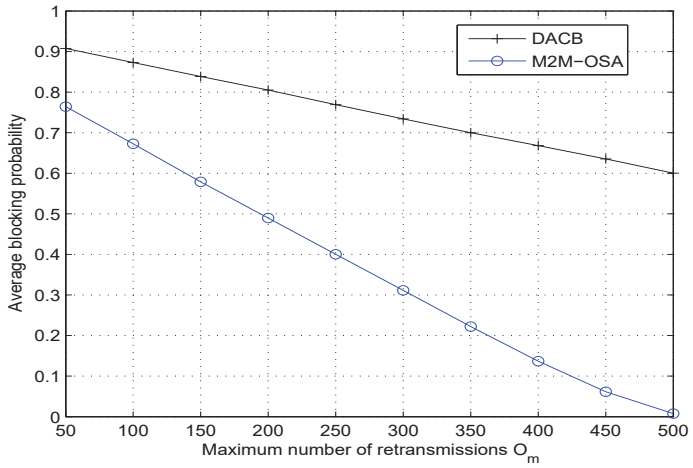

Figure 3.8: Average blocking probability vs. maximum retransmission limit O_m for both M2M-OSA and DACB with $N = 8000$

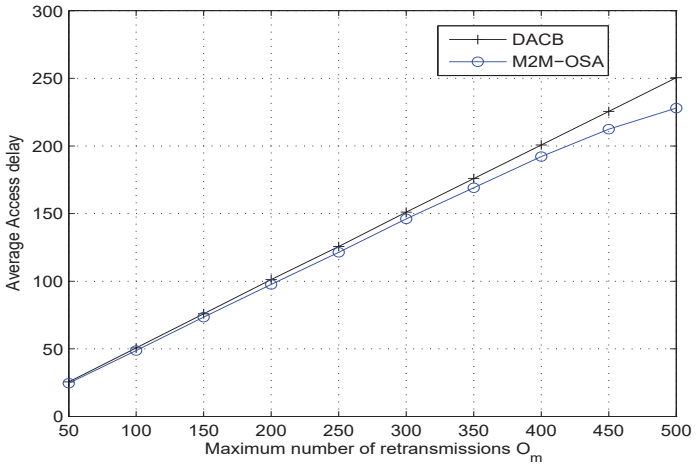

Figure 3.9: Average access delay vs. maximum retransmission limit O_m for both M2M-OSA and DACB with $N = 8000$

as compared to the gap in Figure 3.6. Nevertheless, the superiority of the proposed algorithm still holds since it serves more MTDs, as indicated by Figure 3.8, whereas the decreased access delay in the DACB, achieved by limiting O_m, is a result of the faster access gained by part of the backlogged MTDs due to mitigating congestion by blocking others.

It is worth mentioning that we assume that both the proposed and the DACB techniques have the same number of RBs. Indeed, DACB with the prescribed performance would need more resources for the PRACH in addition to the M RBs used for PUSCH. In particular, these M RBs used for PUSCH should be equal to the sum of expected number of successful and collided preambles. Consequently, sufficient PUSCH resources are available for assignment

to successful preambles and no successful preamble would be blocked due to lack of PUSCH resources [118]. In contrast, M2M-OSA only needs these M RBs for both contention and transmission.

3.7.3 Non-idealities

In this section, we assume that no maximum retransmission limit and that the devices keep trying until gaining successful access. Also, we study the energy consumed by the MTC devices using M2M-OSA which is defined as the total energy spent by an active device until gaining successful access and transmitting its data. In this regard, an active device could be in one of three different states; transmission, reception, or idle state with power consumption $P_t = 50$ mW, $P_r = 50$ mW, or $P_i = 0.025$ mW, respectively. Note that, in the idle case the device is active but not participating in contention, however, it could receive the feedback packets from the BS and rechecks the thresholds for a chance to contend. When a device successes and starts to transmit its data, all other contending devices on the same RB goes to idle state. In the following, the energy is calculated per mini-slot time.

First, we verify the accuracy of the proposed estimation method for the number of active MTDs per time slot. Then, we show the effect of employing this estimate on the reservation time as well as the non-ideal situation of Misestimation. Figure 3.10 plots the number of active devices per time slot for different N. The results show that the estimated N_k is very close to the actual number of devices. The figure also shows that around the 100^{th} time slot, which represents the end of the activation time, the number of backlogged devices reaches its maximum and then starts to decrease while serving the devices with no more new arrivals. Note that the time slots indices at which the curves end represent the total service time. Using the estimated N_k, Figure 3.11 depicts that the average number of mini-slots required for decision, τ, is upper bounded by 2.5 regardless of the massive increase in the number of MTC devices. That is, the proposed algorithm does not introduce much delay for the winner selection even if the actual number of backlogged is unknown.

Figure 3.12 depicts the effect of the misestimation error on the performance of M2M-OSA. Figures 3.12-a, b illustrate that the effect on both average access delay and the TST is minimal, assuming that the time-slot is large enough to accommodate the device's data transmission (small packet in our case). In contrast, the effect is more noticeable on both the average reservation period and the energy as shown in Figures 3.12-c, d. Particularly, for the under-estimation

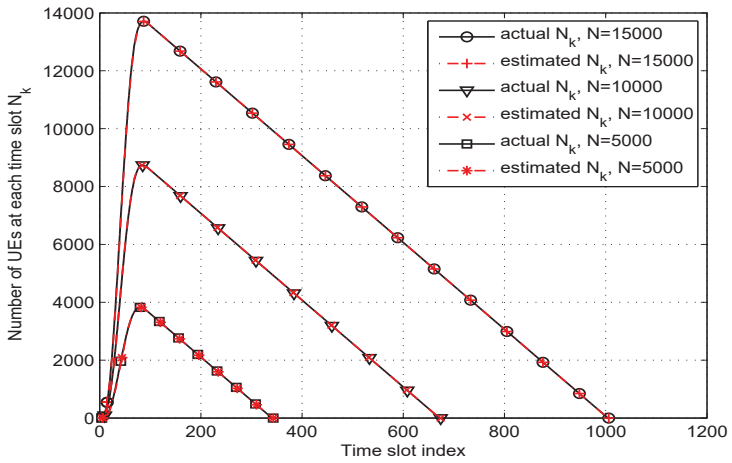

Figure 3.10: A comparison between the actual number of active MTDs per time slot and the estimated one.

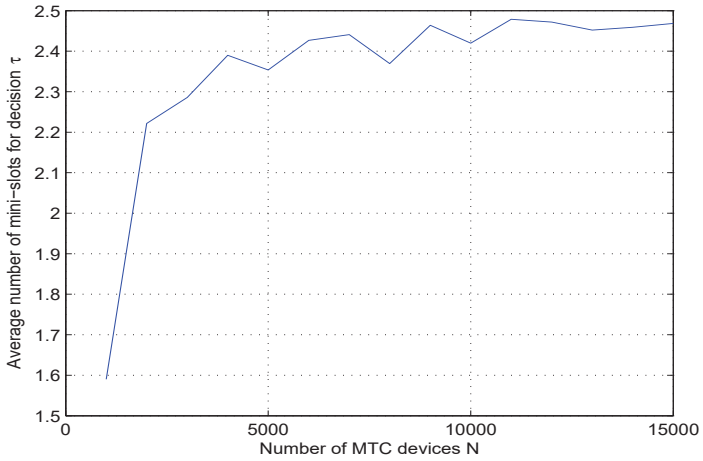

Figure 3.11: Average reservation period in mini-slots unit using M2M-OSA with the estimation of the number of active MTDs

case (with "-" sign), the average contention period is less than the over-estimation case (with "+" sign), whereas the consumed energy is higher. The reason is closely related to the step size as explained before.

For the feedback-error case, the following error rates were considered. First, we assume one error occurs in each $2M$ idle mini-slots and a reinitialization occurs after three consecutive idle mini-slots. Second, we assume one error occurs in each $3M$ success mini-slots. Figures 3.13-a, b depict that the impact of feedback errors on both the averages of access delay and TST is minimal. In contrast, Figures 3.13-c, d show that the errors result in higher average reservation period and energy consumption. A further improvement is to assume that the devices can save the values of the thresholds after the last execution of the "upper" function to jump directly to the

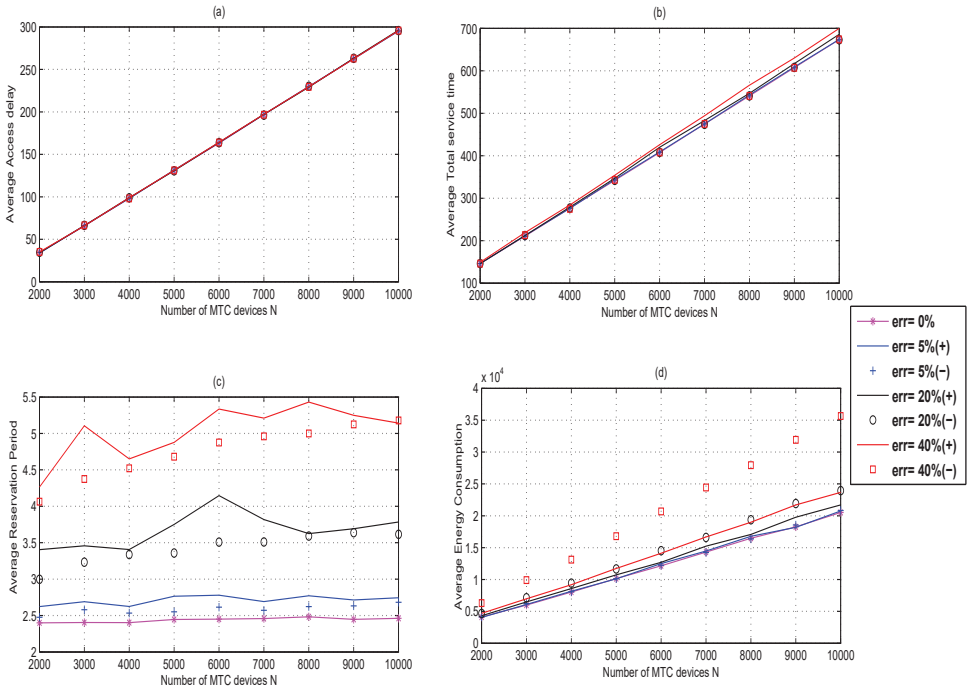

Figure 3.12: Effect of N_{err} on the performance of M2M-OSA

range at which the devices exist if a reinitialization happens due to feedback errors. In general, one can conclude that the errors mainly affect the reservation period which in turn affects the energy consumption. For the case when the time slot is large enough to accommodate the data transmission, both delay and TST will minimally be affected which can be usually the case assuming MTC devices with small amount of data.

For the short time slot scenario, we assume smaller $\zeta = 20$, and add a packet size parameter that indicates the number of mini-slots needed for data packet transmission. Note also that we assume fixed packet size for all devices equals 15. That is if the reservation period exceeds 5 mini-slots (i.e. almost the double of its average), an extension is needed. The effect on both average access delay and TST is shown in Figure 3.14. The figure depicts that the applied modification does not introduce much loss in the performance.

3.8 Summary

The random access (RA) procedure used in Long Term Evolution-Advanced (LTE-A) suffers from frequent collisions, and thus it is inefficient for M2M networks with massive access. In

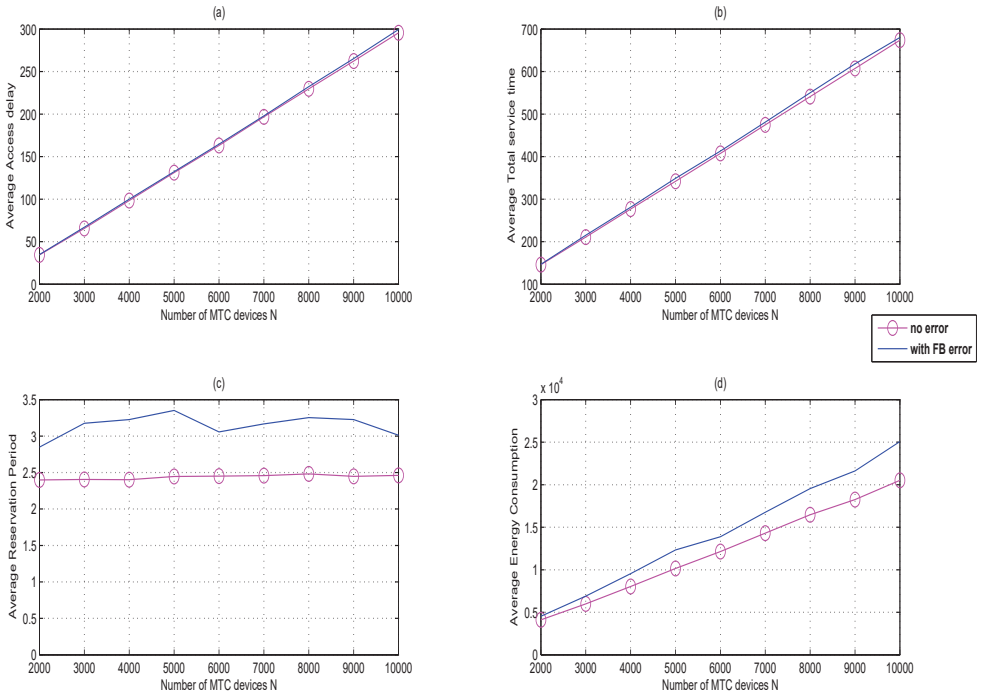

Figure 3.13: Effect of feedback errors on the performance of M2M-OSA

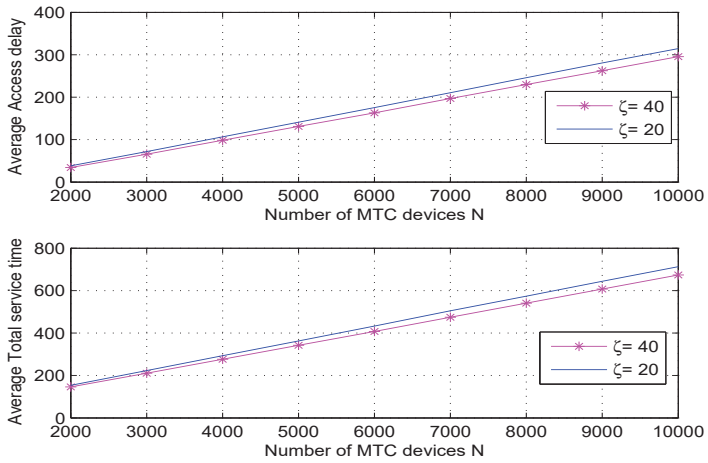

Figure 3.14: Performance evaluation for the scenario where the time slot is short

such networks, network overload problem arises as a performance bottleneck that leads to high resource wastage and severe access delay. In this chapter, we proposed a distributed scalable overload control algorithm for M2M with massive access. The proposed algorithm, M2M-OSA, is able to efficiently handle massive access in M2M networks with full resource utilization and reduced overhead. It was shown that M2M-OSA consumes bounded, relatively small

contention time which is essentially important for MTC with huge number of devices to save energy. We provided a comparison between the proposed M2M-OSA and the existing dynamic ACB (DACB) algorithm with optimal ACB factor. Our simulation results showed that the proposed algorithm outperforms the DACB algorithm in terms of service time, access delay, and blocking probability, especially for dense M2M networks. Furthermore, we provided an estimation method for the number of active devices per time slot including the backlogged ones. This information is important for the proper operation of the proposed algorithm. We also discussed some non-ideal cases such as the effect of misestimation and feedback errors and showed that these errors could be handled using the proposed algorithm with minor modifications while tolerating accepted performance degradation.

4. Performance of Overload Control in Machine-to-Machine Wireless Networks

4.1 Introduction

Machine-to-machine (M2M) communication is an automated data transmission among devices with no or little human intervention. The deployment of M2M networks is expected to enormously increase to support the diverse Internet of Things (IoT) applications. These applications spread over different fields including industrial automation, smart cities and intelligent transportation systems, among others. However, these promising applications come with a number of challenges and the massive access of M2M is a major one.

In literature, the Long Term Evolution (LTE) random access (RA) mechanism was adopted to provide network access for M2M devices. This RA scheme is a 4-step handshaking procedure that begins by a preamble transmission and ends by a contention resolution message from the evolved node B (eNB) if an access success is granted. However, with M2M massive access, this mechanism is rendered a performance bottleneck due to the increased collisions that lead to severe congestion. A very promising scheme, namely access class barring (ACB), was identified by the 3GPP as a solution for M2M overload, among others [6]. Then, several works in literature enhanced its performance such as dynamic ACB (DACB) where the MTDs' access probabilities are adapted dynamically based on the load conditions [8]. However, the improvements attained by this algorithm, and others, are limited in terms of the achieved access delay and resource utilization. Motivated by this, in chapter 3, we proposed a distributed algorithm, referred to as M2M-OSA, that is able to efficiently handle massive access in M2M networks with full resource utilization and bounded delay.

In this chapter, we provide a complete analytical framework to assess the performance the proposed M2M-OSA introduced in Chapter 3. Particularly, a triangular approximation is

adopted to ease the analysis of the Beta-distributed traffic that is known to be mathematically complex. This approximation is then used to find closed-form expressions for the average access delay and total service time considering both lightly, and seriously congested scenarios. Afterwards, the energy consumption is studied such that a per-device average energy consumption expression is provided. For the case of finite retransmission limit, we provide a scheme that enables the calculation of the blocking probability of the devices. In addition, a BIP formulation of the studied problem is provided as a benchmark. The proposed M2M-OSA is then updated to include a traffic regulation mechanism which effectively reduces the energy consumption of the devices. Finally, we shed light on some use cases of the proposed algorithm.

4.2 System Model

In this chapter the system model is the same as the one adopted in Chapter 3. Mainly, we consider the activation of N MTC devices in the coverage area of a single BS according to a Beta probability distribution with parameters $(\alpha = 3, \beta = 4)$ [6]. This activation takes place within bounded activation time T_A. Assume also that the interval $[0, T_A]$ is divided into I_A time slots. Based on that each MTC device is activated at time $t \in [0, T_A]$ with probability $g(t)$ calculated as follows:

$$g(t) = \frac{t^{\alpha-1}(T_A - t)^{\beta-1}}{T_A^{\alpha+\beta-1}\mathcal{B}(\alpha, \beta)}, \tag{4.1}$$

where $\mathcal{B}(\alpha, \beta)$ is the beta function [114]. For simplicity, we assume that a device is activated at the beginning of the time slot and performs its first access attempt during the same time slot [8]. We also assume M uplink resources (referred to as resource blocks (RBs)) that are exclusively available for M2M use. Hence, no collision between human-to-human (H2H) traffic and M2M traffic. Furthermore, making use of the small-sized packets generated from M2M devices [3], we adopt fast data transmission right after access is granted [56].

In the time domain, we assume that a time slot consists of ζ mini-slots. Additionally, the duration of a mini-slot is assumed to be enough to accommodate the transmission of a small request packet and receive its feedback. Mainly, the time slot consists of two different periods, reservation and transmission. During the reservation mini-slots, the devices contend by transmitting request packets to the BS. The successful device then is able to transmit its data during the transmission period that is represented by the remaining mini-slots till the time slot ends. Hence, it is essentially important to keep the number of reservation mini-slots as low as

possible to save more time for useful data transmission. Furthermore, a successful access is considered as long as a winner MTD is selected whereas the physical channel impairments are neglected [53].

4.3 Performance Analysis

In this section, we analyze the performance of the proposed M2M-OSA in terms of the average access delay, total service time (TST), energy consumption and blocking probability.

4.3.1 Infinite Retransmission Limit

Here, we study the case of $O_m \rightarrow \infty$ where no devices are blocked.

4.3.1.1 Average Access Delay

The access delay of an MTC device is defined as the number of time slots elapsed from the first access attempt until gaining successful access. As suggested by the 3GPP, Beta distribution is used to model the arrival of MTC devices in overload situations, and so it is adopted in our proposal. However, the algebraic analysis of the Beta distribution is known to be difficult [119]. To study the M2M traffic, following [120], a triangular approximation to the Beta distribution that eases its mathematical tractability is used. This approximation was also employed to find closed-form expressions for the access delay resulted from DACB [53]. Particularly, the obtained expressions were considered as lower-bounds since an optimal ACB factor was assumed. In the following, we analytically find the access delay for the proposed M2M-OSA.

Define the channel capacity, δ, as the maximum rate of successes achieved by the devices in the network. The access delay is analyzed in two different cases. In the first case, the maximum arrival rate, λ_{max}, never exceeds the channel capacity, δ, and is referred to as stable backlog. In the second case, the network is seriously congested where the maximum arrival rate significantly exceeds the channel capacity, which we refer to as unstable backlog. This scenario results in accumulated backlog that is eliminated after the activation interval ends.

The following shows the triangular approximation to the Beta distribution. In particular, the

triangular density function and the average arrival rate are given by (4.2), (4.3), respectively.

$$f(x;a,b,m) = \begin{cases} \frac{2}{b-a}\frac{x-a}{m-a} & a \le x \le m, \\ \frac{2}{b-a}\frac{b-x}{b-m} & m \le x \le b, \end{cases} \tag{4.2}$$

$$\lambda(t;a,b,m) = \begin{cases} \frac{N}{T_A}\frac{2}{b-a}\frac{t/T_A-a}{m-a} & a \le t/T_A \le m, \\ \frac{N}{T_A}\frac{2}{b-a}\frac{b-t/T_A}{b-m} & m \le t/T_A \le b, \end{cases} \tag{4.3}$$

where a, b, m are the parameters of the triangular approximation [121]. From (3.8), the average number of successes can be approximated by:

$$\bar{S}_k \approx M(1-e^{-N_k/M}). \tag{4.4}$$

In case of congestion where $N_k >> M$, unlike in RA procedure where collisions increase, M2M-OSA drives the system towards full resource utilization. Therefore, by ignoring the exponential term in (4.4), the channel capacity is defined as:

$$\delta = M. \tag{4.5}$$

For the stable backlog case, using the approximation in (4.3) and the result of (4.5), the stability condition, $\lambda_{max} < \delta$ can be rewritten as:

$$N < \frac{MI_A(b-a)}{2}. \tag{4.6}$$

Note that T_A is replaced by I_A since we study the network on time-slot basis. In the stable backlog case, the average number of successes is equal to the average arrivals at each time slot. The reason is that the arrivals are less than the capacity, so in average all arrivals should succeed. However, in reality some of the arrivals could fail so they are backlogged added to the arrivals of the next slot. Therefore, having a constant of x arrivals per time slot, the backlog converges to $y(x) > x$, and the average number of successes is also x. Based on that, one can represent (4.4) as follows:

$$\begin{aligned} M(1-e^{-y(x)/M}) &= x \\ y(x) &= -M\ln\left(1-\frac{x}{M}\right). \end{aligned} \tag{4.7}$$

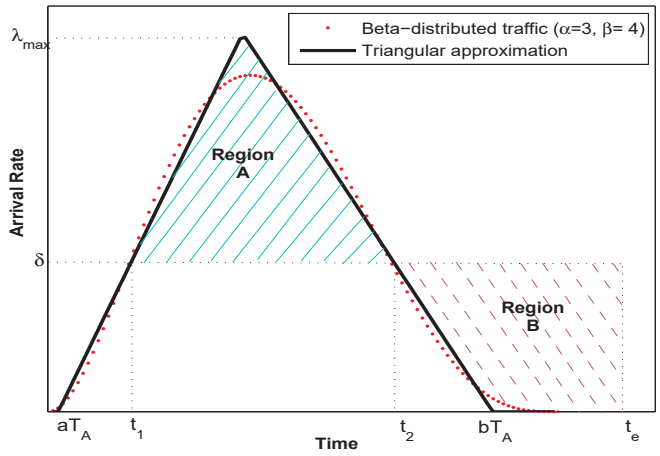

Figure 4.1: Triangular approximation for the Beta distribution in unstable backlog case ($a = 0.0208, b = 0.879, m = 0.385$)

For Beta-distributed traffic, neither x nor $y(x)$ is a constant. However, for simplicity it can be assumed that, for a certain device, the number of backlogged devices is constant within the time slots at which the device attempts to get access [53]. Based on that, and by defining $L = \lambda_{max}/M$, the average access delay for the stable case, d_s, can be calculated as follows:

$$\begin{aligned} d_s &= \frac{\int_0^{\lambda_{max}} y(n)\mathrm{d}n}{\int_0^{\lambda_{max}} n\mathrm{d}n} \\ &= \frac{2}{L^2}\left((1-L)\ln(1-L)+L\right), \qquad \text{if } N < 0.5MI_A(b-a). \end{aligned} \tag{4.8}$$

For the case of highly congested networks, the delay expression in [53] follows directly from studying the triangular approximation. Hence, the provided expression works also in our M2M-OSA while using our own $\delta = M$. Figure 4.1 shows the triangular approximation of the Beta distribution. On the approximated triangular, the arrival rate of the devices is larger than δ between times t_1 and t_2 where its peak value, $\lambda_{max} = \frac{N}{T_A}\frac{2}{b-a}$, significantly exceeds δ. Thus, the accumulated backlog is completely eliminated after bT_A at time t_e. By examining the geometry, the times t_1 and t_2 could be obtained using (4.3) whereas $\lambda(t) = \delta$ as follows:

$$\begin{aligned} t_1 &= \frac{MT_A{}^2(a-b)(a-m)}{2N} + aT_A, \\ t_2 &= \frac{MT_A{}^2(a-b)(b-m)}{2N} + bT_A. \end{aligned}$$

Since the integral of the arrival rate gives the number of nodes, in Figure 4.1, the area of re-

gion A represents the number of accumulations between t_1 and t_2, whereas the area of region B represents the amount of successes from this accumulated backlog achieved after t_2. The end time, t_e, is obtained from equating the two areas; regions A and B in Figure 4.1 as follows:

$$\frac{1}{2}(\lambda_{max} - \delta)(t_2 - t_1) = \frac{1}{2}(bT_A - t_2)\delta + (t_e - bT_A)\delta$$

This end time, t_e, is known formally as the total service time (TST) which is the total time needed for all devices to successfully access the channel. In a time-slot basis, the TST, k_e, is calculated approximately as follows [53]:

$$k_e = \frac{MI_A^2(a-b)(a-m)}{4N} + aI_A + \frac{N}{M}. \tag{4.9}$$

It shall be emphasized that k_e is the discrete version of t_e that counts the total number of time slots needed for all devices to get successful access. The expression in (4.9) is only valid for $k_e > bI_A$. By neglecting the delay experienced by the devices arriving before t_1 and using the end time, [53] obtained the mean arrival and departure times in the network. Getting the difference between these two times, the average access delay for the unstable backlog case, d_u, is given by:

$$\begin{aligned} d_u = &\Big(32MN^3I_A(2a-b-m) + 24M^2N^2I_A^2(a-b)(a-m) \\ &- M^4I_A^4(a-b)^2(a-m)^2 + 48N^4\Big)/\Big(96MN^3\Big) + 1, \\ &\text{if } N > 0.5MI_A\left(\sqrt{(b-a)(b-m)} - a + b\right). \end{aligned} \tag{4.10}$$

As mentioned earlier, the condition forced for d_u calculation is to ensure the serious congestion situation such that the total service time is larger than the activation time.

4.3.1.2 Energy Consumption

Here, we study the energy consumed by the MTC devices using M2M-OSA. The energy consumption is defined as the total energy spent by an active device until gaining successful access. In this regard, an active device could be in one of three different states as follows:

- Transmission (Tx) state with power consumption P_t
- Reception (Rx) state with power consumption P_r

- Idle state with power consumption P_i

We provide an expression to analytically calculate the average energy consumption under certain assumptions. First, we use the upper bound of 2.507 mini-slots for the average reservation period in the calculation. Let, T_{Ms} is a mini-slot time, and $T = \zeta T_{Ms}$ is the time of one time slot, hence, the expected reservation period is $E\{T_{res}\} = 2.507 T_{Ms} = \frac{2.507}{\zeta} T$. Second, we assume that this reservation period is divided as follows; one success mini-slot, and the remaining reservation time is equally divided between collision and idle events. This is valid under the assumption that both collision and idle states are equiprobable for the sake of simplicity. Additionally, we adopted the assumption that most likely a collision event involves two devices [83, 117]. These assumptions are also validated by the accuracy of the simulation results. Based on that, and by using the expected number of both devices, $\bar{N}_k$, and successes, $\bar{S}_k$, at each time slot, the average energy consumption for both stable and unstable backlog cases are calculated.

First, for the reservation period, the energy consumption during collisions, E_{R_c}, is given by:

$$E_{R_c} = \frac{E\{T_{res}\} - 1}{2} \left(2\bar{S}_k \frac{P_t + P_r}{2} + (\bar{N}_k - 2\bar{S}_k) \frac{P_i + P_r}{2} \right) \tag{4.11}$$

where $(P_t + P_r)/2$ and $(P_i + P_r)/2$ represent the energy spent by the devices per mini-slot if sending transmission requests or not, respectively. Also, $\bar{S}_k$ is used to represent the average number of active RBs where a contention is running. Note that, by dropping T_{Ms} from the energy equations, the energy consumption is obtained per mini-slot time.

In the reservation period, an idle feedback is announced by the BS when none of the devices transmit request. Thus, the energy consumed in an idle-resulted contention period, E_{R_i}, is:

$$E_{R_i} = \frac{E\{T_{res}\} - 1}{2} \left(\bar{N}_k \frac{P_i + P_r}{2} \right). \tag{4.12}$$

In a success mini-slot with one transmitting device per RB, the average consumed energy, E_{R_s}, is:

$$E_{R_s} = 1 \times \left(\bar{S}_k \frac{P_t + P_r}{2} + (\bar{N}_k - \bar{S}_k) \frac{P_i + P_r}{2} \right). \tag{4.13}$$

Given the above, the average consumed energy for reservation at all RBs per time slot, E_R, is:

$$E_R = E_{R_c} + E_{R_i} + E_{R_s}. \tag{4.14}$$

Next, we calculate the average energy consumed during the transmission period of a time slot, E_T. During the transmission period, all the devices except the succeeded ones (i.e. $\bar{N}_k - \bar{S}_k$), are at the idle state with power consumption P_i. Thus, the energy E_T is given by:

$$E_T = (\zeta - E\{T_{res}\}) \left(\bar{S}_k P_t + (\bar{N}_k - \bar{S}_k) P_i\right). \tag{4.15}$$

To that end, the average energy consumed by the devices in the system at the k^{th} time slot is $E_k = E_R + E_T$. Therefore, the average energy per device is obtained as follows:

$$E_{avg} = \frac{1}{N} \sum_{k=1}^{k_e} E_k. \tag{4.16}$$

4.3.2 Finite Retransmission Limit

Here, we study the case of finite O_m where the blockage of some devices is possible. We analyze two important metrics; namely, the blocking probability and the average access delay.

4.3.2.1 Blocking Probability

The blocking probability, P_b, is the ratio between the number of blocked MTC devices due to exceeding their maximum retransmission limit, O_m, to the total number of MTC devices registered in the network, N. In the following, we propose a method that statistically obtains the blocking probability resulted from M2M-OSA. At each time slot, the number of contending devices is the sum of the new arrivals and the backlogged devices from previous slots as illustrated in (3.4). However, if the network enforces a finite O_m, then N_k is also affected by the number of blocked devices, b_k, as follows:

$$N_k = \begin{cases} N_{k-1} - S_{k-1} + a_k - b_k & \text{if } k \leq I_A, \\ N_{k-1} - S_{k-1} - b_k & \text{otherwise.} \end{cases}$$

In order to obtain the blocking probability, P_b, an estimate for b_k is needed. Thus, the number of attempts by each device should be tracked. To that end, we adopt another formulation for N_k that distinguishes between different *classes* of backlog as follows [122]:

$$N_k = \sum_{j=1}^{O_m} N_k[j] \tag{4.17}$$

where $N_k[j]$ is the number of devices at their j^{th} attempt in the k^{th} time slot, and it could be considered as the j^{th} backlog class. For instance, $N_k[1] = a_k$. To be able to track the values of $N_k[j]$ of different classes, one needs to know the share of each class on the total number of successes at each time slot. This can be done statistically as follows:

$$\bar{N}_k[1] = \begin{cases} \lambda_k & \text{if } k \leq I_A, \\ 0 & \text{otherwise,} \end{cases} \tag{4.18}$$

$$\bar{N}_k[j] = \bar{N}_{k-1}[j-1] - \frac{\bar{N}_{k-1}[j-1]}{\bar{N}_{k-1}} \bar{S}_{k-1}, \quad j = 2, 3, \ldots, O_m. \tag{4.19}$$

Note that, in M2M-OSA, all active devices sharing the same RB have the same opportunity of success regardless of the number of attempts. This is because, the winner selection is based on a contention metric randomly selected from the uniform distribution at each time slot[1]. Hence, the second term of (4.19) describes the share of class $(j-1)$ in the number of successes by simply dividing the number of devices at this class by the total number of active devices, in the average sense. Using $\bar{N}_k$, the average number of successes at the k^{th} time slot, $\bar{S}_k$, can be evaluated from (3.8).

Given the above, the average number of blocked devices can be calculated by counting the devices of the $(O_m + 1)^{th}$ class at each time slot as follows:

$$\begin{aligned} \bar{N}_k[O_m + 1] &= \sum_{i=1}^{k} \bar{b}_i \\ &= \bar{N}_{k-1}[O_m + 1] + \bar{N}_{k-1}[O_m] - \frac{\bar{N}_{k-1}[O_m]}{\bar{N}_{k-1}} \bar{S}_{k-1} \end{aligned} \tag{4.20}$$

Finally, P_b can be calculated as:

$$P_b = \frac{\bar{N}_{k_e}[O_m + 1]}{N} \tag{4.21}$$

where the subscript k_e refers to the end time slot (i.e TST) where all devices either gained successful access or blocked. Algorithm 3 shows the procedure to calculate the average P_b and k_e.

Table 4.1 is a numerical example illustrating the procedure of Algorithm 3. In this example, we assumed that $M = 3$, $O_m = 3$. Additionally, for simplicity, we directly assumed values

[1]The important feature is that the contention metric changes randomly at each time slot, so any user can meet the selection criterion at any attempt. Thus, the MTDs' equiprobable success opportunities resulted from randomness not the uniform distribution.

Algorithm 3 Procedure for getting P_b and k_e of M2M-OSA

Given N, M, I_A, O_m
Calculate λ vector from (3.6), $\lambda = [\lambda_1\ \lambda_2\ \ldots\ \lambda_{I_A}]$
Initialize: $k = 1, \bar{N}_k = \bar{N}_k[1] = \lambda_k, \sum_{j=2}^{O_m+1} \bar{N}_k[j] = 0$
while $\bar{N}_k \geq 0$ **do**
  $\bar{N}_k = \sum_{j=1}^{O_m} \bar{N}_k[j]$
  Use (3.8) to get $\bar{S}_k$
  $k = k + 1$
  Use (4.18), (4.19) to get $\bar{N}_k[j]$
  Use (4.20) to track the number of blocked devices
end while
Use (4.21) to get P_b
$k_e = k$

Table 4.1: Example of P_b Calculation Procedure

k	$\bar{N}_k[1]$	$\bar{N}_k[2]$	$\bar{N}_k[3]$	$\bar{N}_k[4]$	$\bar{N}_k$	$\bar{S}_k$
1	15	0	0	0	15	3
2	18	12	0	0	30	3
3	21	16.2	10.8	0	48	3
4	24	19.68	15.18	10.125	58.87	3
5	27	22.77	18.68	24.53	68.46	3

for a_k's; however, their actual values depend on the Beta distribution parameters and the values of N and I_A. As depicted in Table 4.1, for example, the value of $\bar{N}_3[2] = 16.2$ is obtained using (4.19) such that $\bar{N}_3[2] = 18 - \frac{18\times3}{30} = 16.2$, and similarly the other table entries could be obtained. In this example, $\bar{N}_k[4]$ is the total number of blocked MTDs till the k^{th} cycle, and it can be used to calculate P_b at the end. Note that, Algorithm 3 can also help in estimating the number of MTDs at each time slot when O_m is finite (i.e. column 6 in Table 4.1). In particular, given certain system parameters, the BS can run the algorithm off-line and broadcasts the value of $\bar{N}_k$ at each time slot to all MTDs.

4.3.2.2 Average Access Delay

For the case of $\lambda_{max} > M$, define k_1 as the first time slot such that $\lambda_{k_1} > M$, and that t_1 is its start time. By ignoring the access delay for devices before k_1, the expected arrival and departure

times of devices at and after k_1 are given respectively by:

$$\begin{aligned} E\{t_{arr}\} &= \left(\int_{t_1}^{T_A} \lambda(t) t \mathrm{d}t\right) / \left(\int_{t_1}^{T_A} \lambda(t) \mathrm{d}t\right) \\ &= \Big(- M^3 I_A^4 (a-b)^2 (a-m)^2 - 3aM^2 N I_A^3 (a-b) \times (a-m) \\ &\quad + 4N^3 I_A (a+b+m)\Big) / \Big(12N^3 - 3M^2 N I_A^2 (a-b)(a-m)\Big). \end{aligned} \tag{4.22}$$

$$E\{t_{dep}\} = (k_e + k_1)/2. \tag{4.23}$$

Let $\bar{N}_{[k_1, I_A]}$ be the average number of devices arriving at and after k_1 which can be obtained from the λ vector as illustrated in Algorithm 3. Then, the average access delay is calculated as follows:

$$d = \left(E\{t_{dep}\} - E\{t_{arr}\}\right) \frac{\bar{N}_{[k_1, I_A]}}{N}. \tag{4.24}$$

For the stable backlog case, from (4.8), it can be shown that at the edge value where $\lambda_{max} = M$, $\lim_{L \to 1} d_s = 2$. This means that the maximum average access delay is relatively small. Therefore, for relatively large O_m, $P_b \to 0$ and the expression in (4.8) is valid. However, in unlikely situations where relatively small O_m is used, a few devices could be blocked which turns (4.8) invalid. One approximate solution is to find a virtual value N_v that represents the number of devices succeeded in the system, and use N_v instead of actual N in (4.8). This approximate solution could be accepted since the deviation from the correct value is a small fraction, hence the error will be small.

4.4 Binary Integer Programming Formulation

In this section, a BIP problem is formulated to compare the performance of the proposed algorithm to the optimal one in terms of the average access delay. Although BIP problems are solvable for small number of variables, this formulation helps to assess the performance of the proposed algorithm. The studied problem can be formulated as a BIP problem as follows:

$$\max \quad \sum_{j=1}^{M} \sum_{n=1}^{N_k} w_{n,j} y_{n,j} \tag{4.25}$$

$$\text{s.t.} \quad \text{C1: } \sum_{n=1}^{N_k} y_{n,j} \leq 1, \quad \forall j \tag{4.26}$$

$$\text{C2: } \sum_{j=1}^{M} y_{n,j} \leq 1, \quad \forall n \tag{4.27}$$

$$\text{C3: } y_{n,j} \in \{0, 1\}, \quad \forall n, j \tag{4.28}$$

where $y_{n,j}$ is a scheduling binary variable that takes the value 1 when the n^{th} device is scheduled on the j^{th} RB and zero otherwise. The objective of the optimization problem is to maximize the metric O. This is done by setting the weights $w_{n,j}$ as follows:

$$w_{n,j} = O_n p_{n,j}$$

where $p_{n,j}$ is the preference value of the n^{th} device on the j^{th} RB. The preference values are arbitrary values given by the devices to each RB to represent the devices' preference. As $p_{n,j}$ increases, this means that device n prefers the j^{th} RB more. Note that, maximizing O allows for the selection of the devices that are delayed most in the system and so their priorities increase. This selection criterion results in reducing the average access delay of the system. For the conditions, C1 indicates that each RB cannot be allocated to more than one device, whereas C2 assures that each device is allocated no more than one RB. By adapting the preference values, we allow each device to randomly select one RB by putting all its preference budget on a single RB and giving zero to all other RBs. This is to mimic the step of randomly selecting single contention RB in the M2M-OSA case. The problem was solved using Matlab.

4.5 M2M-OSA with Traffic Regulation

Inspired by the fact that no devices more than the number of RBs could succeed at each time slot, it is inefficient to keep a massive number of devices active and competing at each time slot. Applying a good traffic regulation mechanism would save considerable amount of energy while taking into account not to affect the system utilization (no RB is wasted) or the devices' average access delay. Particularly, at the end of the activation period the BS uses the estimate of the number of backlogged devices in the system to announce a maximum backoff value B calculated as follows:

$$B = \frac{\bar{N}_{I_A+1}}{f * M} \tag{4.29}$$

where f is a traffic factor defined by the system. The idea behind the control variable f is to guarantee an average of f devices per RB such that no RB is wasted due to the lack of the competing devices. Afterwards, each device n draws a random number d_n from the uniform distribution $U[1, B]$. The selected number is the index of the time slot at which the device will be activated for the second time. Note that, this index is shifted by an offset I_A since the traffic

regulation starts when the activation period ends. Thus, the n^{th} device sleeps $d_n - 1$ time slots and saves the energy consumed in competition at these slots.

According to this procedure, the backlogged devices that remain in the system after I_A time slots enters a second round of activation that consumes B time slots. But, on this round, the activation follows a uniform distribution not a Beta one as before. Hence, the average number of newly activated devices at each time slot k equals:

$$\alpha_k = \begin{cases} \frac{\bar{N}_{I_A+1}}{B} & , k = I_A + 1, \ldots, I_A + B, \\ 0 & , \text{otherwise}. \end{cases} \tag{4.30}$$

Since, the system is in a fresh-start state, the devices reset the number of backlogged to zero. Therefore, the average number of devices is:

$$\bar{N}_k = \alpha_k \qquad , k = I_A + 1 \tag{4.31}$$

$$\bar{N}_k = \begin{cases} \bar{N}_{k-1} - \bar{S}_{k-1} + \alpha_k & , k = I_A + 2, \ldots, I_A + B, \\ \bar{N}_{k-1} - \bar{S}_{k-1} & , k > I_A + B. \end{cases} \tag{4.32}$$

4.6 Simulations and Discussions

We consider a single cell scenario while varying the number of MTC devices N up to 15000 devices that are all cell-synchronized. A Beta-distributed traffic model is simulated with ($\alpha = 3, \beta = 4$), and we consider a device is active if it has data to be transmitted. As before, the proposed algorithm is compared to the DACB scheme with optimal ACB factor AP [107]. In DACB, the BS dynamically broadcasts the optimal $AP = \min(1, M/N_k)$ which depends on the *actual* number of backlogged devices. Only the devices whose random draws are less than AP will perform the 4-step RA procedure. Table 4.2 shows the simulation parameters. We examine the results in both cases of O_m, infinite and finite.

4.6.1 Infinite Retransmission Limit

Figure 4.2 shows the average access delay considering infinite O_m. Particularly, Figure 4.2-a shows the results for the stable backlog using (4.8) and simulations for both actual and estimated N_k. Note that the average access delay is measured in terms of the number of consumed

Table 4.2: System model parameters

Parameter	Definition	Value
M	Number of uplink resources (RBs)	15
N	Number of MTC devices	Stable: 100:100:600 Unstable: 2000:1000:15000
I_A	Number of slots per activation period	100
ζ	Number of mini-slots per time slot	40
a	Parameter for triangular approx.	0.0208
b	Parameter for triangular approx.	0.879
m	Parameter for triangular approx.	0.385
P_t	Transmission power	50 mW
P_r	Reception power	50 mW
P_i	Idle power	0.025 mW

time slots by each device until gaining access. As shown from the results, our approximation and the simulation results are in very good agreement. Also, the results show that the access delay increases with the increase of N due to the increased contention. For unstable backlog, Figure 4.2-b first compares the analysis and the simulation where it shows almost similar performance. Thus, we only considered the simulation curve with $\bar{N}_k$. Second, the figure compares the performance of both M2M-OSA and DACB scheme with optimal ACB factor. As shown, M2M-OSA achieves high improvement in the access delay that is more noticeable for seriously congested scenarios. This is due to the fact that M2M-OSA does not suffer from the increased collisions resulted from the RA procedure. In contrast, the M2M-OSA in seriously congested cases is able to achieve full resource utilization. It can be noticed that the performance gap between the two techniques increases with the increase of N. Consequently, the proposed algorithm provides better adaptation to network scalability, whereas DACB suffers from increased collisions and resource wastage. It is worthy to mention that decreasing the average access delay has a direct impact on reducing the energy consumption which is a key requirement for battery-constrained M2M communications.

By examining the TST results, one can verify that M2M-OSA drives the congested system to full resource utilization. Figure 4.3 plots the TST for unstable backlog for both M2M-OSA and DACB. Additionally we plot the maximum-utilization curve that represents no wasted resources, simply by calculating N/M. First, the figure depicts that M2M-OSA closely approaches the maximum-utilization result, and achieves much better performance than DACB. Indeed, DACB consumes more time slots to compensate for the wasted resources in collisions. Finally, the results show the accuracy of the analytical results in (4.9) compared to the sim-

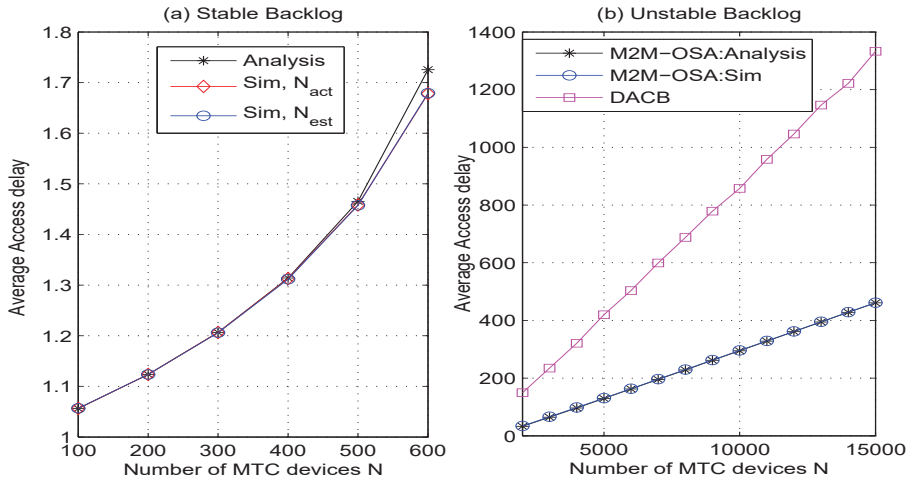

Figure 4.2: Average access delay vs. number of devices with no maximum retransmission limit

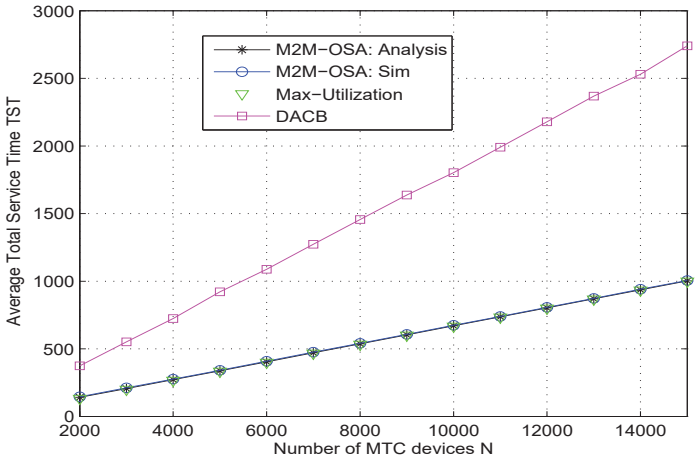

Figure 4.3: Average total service time (i.e. k_e) for unstable backlog with no maximum retransmissions limit.

ulation results. More importantly, the results in this figure endorse that M2M-OSA supports network scalability.

To further prove the effectiveness of M2M-OSA, Table 4.3 shows the average access delay achieved by the proposed M2M-OSA and that resulted from solving the BIP problem using MATLAB for different values of N. The table shows that the results of both schemes are almost the same. Hence, the simple procedure of just drawing a random number from the uniform distribution is enough to minimize the average access delay. This result is confirmed by adding the results of a centralized algorithm, referred to as MAX-O. In MAX-O, the winner device at each RB is selected by the BS in a centralized manner based on its overload indicator O. This winner is the one with the highest O at each RB. Thus, each device should feedback

Table 4.3: Average access delay resulted from different techniques

N / Technique	**200**	**500**	**800**	**1000**
M2M-OSA (actual N_k)	1.1216	1.4585	2.8544	6.4373
M2M-OSA (estimated N_k)	1.1213	1.4606	2.8550	6.4448
M2M-BIP	1.1210	1.4552	2.8597	6.4327
MAX-O	1.1260	1.4552	2.8663	6.4319
Optimal DACB	1.4460	14.1052	39.0710	57.2973

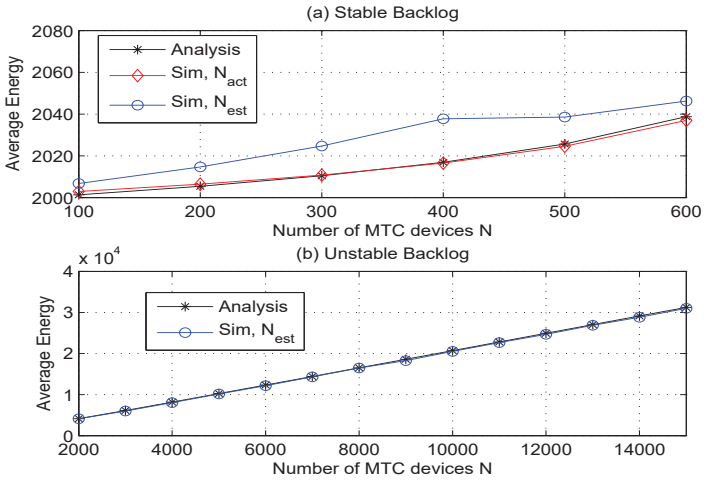

Figure 4.4: Average Energy vs. number of devices with no maximum retransmission limit

its O to the BS which results in increased communication overhead. Also, for fair comparison, it is assumed in MAX-O that each device is allowed to randomly select a single RB to compete for as in M2M-OSA. This assumption reduces the overhead when employing the M2M-OSA since each device transmits its request packet on one RB only. Moreover, this is similar to the conventional random access procedure where each device randomly selects one preamble. The DACB with optimal factor results are also added to show the large improvement achieved by the proposed algorithm.

Regarding the energy consumption, Figure 4.4-a compares the simulation and analysis results for the stable backlog case, whereas Figure 4.4-b illustrates the unstable backlog case (i.e. seriously congested case). The results show the accuracy of our analytical results compared to simulations. One can notice that, the effect of using $\bar{N}_k$ is more noticeable in energy calculations than in average access delay. This is due to the fact that the energy is very sensitive to the status of each single mini-slot whether contention (idle, success, or collision) or transmission, whereas the access delay is obtained on time-slot basis. More precisely, the estimate of the number of backlogged devices affects the threshold updating step in OSA, and hence the

Table 4.4: Validation results for the energy calculation assumptions

N	P(collision)	P(idle)	Avg. devices involved in a collision
2000	0.4415	0.5585	2.2722
5000	0.4613	0.5387	2.2770
15000	0.4666	0.5334	2.2823

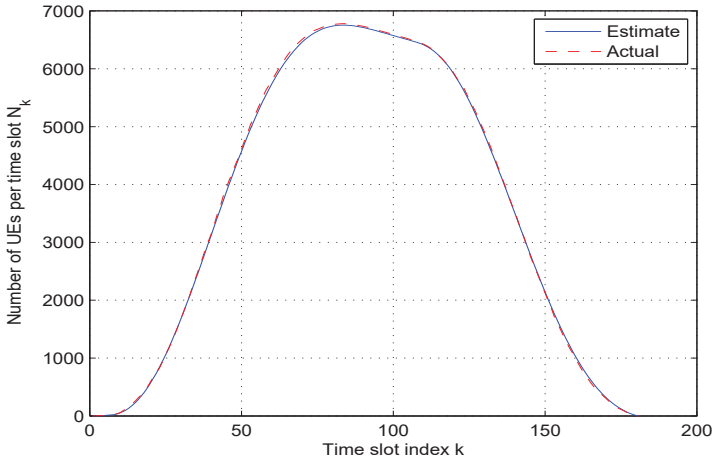

Figure 4.5: A comparison between the number of UEs at each time slot actual vs. estimated for $N = 8000, O_m = 100$.

number of reservation mini-slots. However, since $\bar{N}_k$ is the average of N_k, it still satisfies the 2.5 upper bound for the average reservation period while it may sometimes achieve higher τ than the case of actual value. Table 4.4 shows the results that validate the assumptions adopted while analytically calculating the average energy consumption. The probabilities of both idle and collision events are calculated by dividing the number of occurrence of each event by the sum of both events.

4.6.2 Finite Retransmission Limit

Here, we discuss the results of the case when O_m is finite. Particularly, the simulation results are compared to the analysis obtained using Algorithm 3. As mentioned earlier, Algorithm 3 can help in estimating N_k as Figure 4.5 depicts for $N = 8000, O_m = 100$; however, the actual N_k is used to generate the following results.

By increasing N up to 30000, we examine the average blocking probability in Figure 4.6. The results validate the effectiveness of Algorithm 3 by showing that the analysis is very close to the simulation while changing different system parameters N, M, O_m. The figure clarifies the effect of increasing the available resources M in alleviating congestion and decreasing the prob-

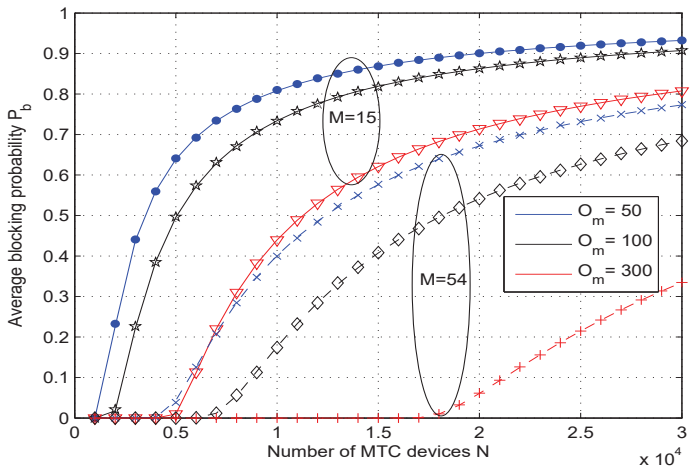

Figure 4.6: Average blocking probability analysis vs. simulations for different number of UEs N, while changing O_m, M. (lines: analysis, markers: sim)

ability of blocking. For the sake of comparison, Figure 4.7 plots P_b and the average access delay for both M2M-OSA and DACB with $N = 8000$ and different values of O_m. Figure 4.7-a shows that the M2M-OSA achieves enhanced P_b as compared to DACB, and the gap in performance increases as O_m increases. For high O_m values, the probability of congestion increases since the MTDs can persist. Therefore, the collisions increase and more resources are wasted using DACB as a result of RA. On the other hand, although with high O_m more MTDs get involved in contention, their chance of getting success increases by allowing more attempts which leads to an improvement in the overall blocking probability of the system. However, this scenario will obviously increase the average access delay. Figure 4.7-b depicts a little enhancement in the average access delay resulted from M2M-OSA as compared to DACB. However, this result should be combined with the P_b result to see the full picture. In particular, the calculation of the average access delay only accounts for the succeeded MTDs. So that, the access delay resulted from M2M-OSA is parallelized with serving high number of MTDs. In contrast, the less succeeded MTDs in the DACB system were able to obtain faster access since congestion is mitigated by blocking others.

Figure 4.8 shows the accuracy of our analytical results in (4.24) for the average access delay compared to the simulations for the case of finite O_m. Additionally, the figure shows that the average access delay starts to increase with N but finally saturates regardless of the increase of N. Indeed, the results reveal that the saturation limit has a strong relation with O_m. While changing O_m, Figure 4.8 shows that the average access delay saturates around half O_m. One can interpret this by focusing on a case of large N where congestion is serious as in Figure 4.9.

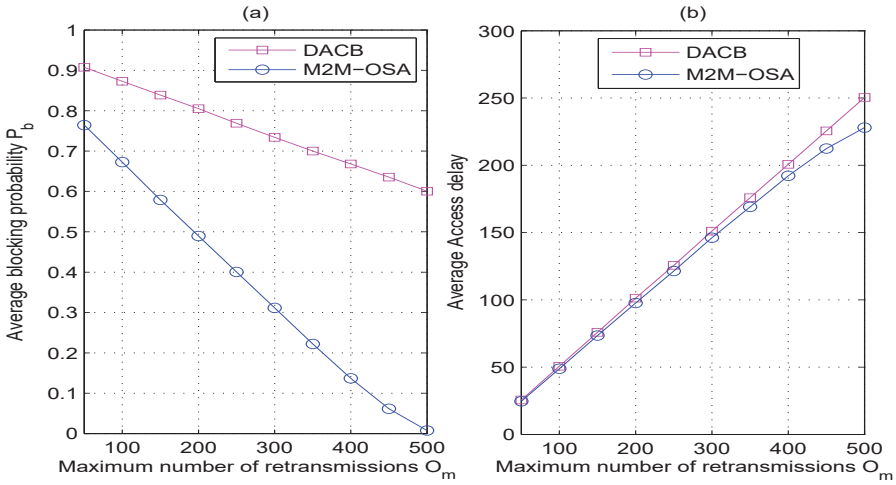

Figure 4.7: A comparison between M2M-OSA and DACB with $N = 8000$ and different O_m. (a) Average blocking probability (b) Average access delay

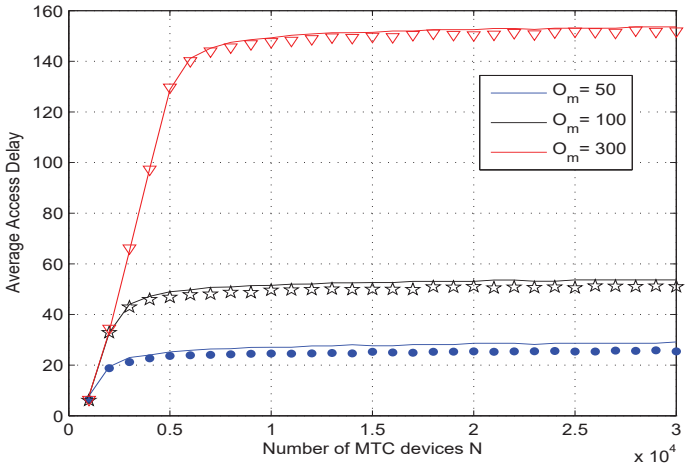

Figure 4.8: Average access delay analysis vs. simulations for different number of UEs N and different O_m values. (lines: analysis, markers: sim)

The figure is a histogram for the access delay with $N = 20000$, and $O_m = 300$. The figure illustrates the frequency of achieving each access delay value by the devices in the system. The results show that, excluding the edges, the access delay approximately follows a uniform distribution with mean $O_m/2$. In such a case, at a congested time slot k (e.g. at region B), N_k is described by (4.17) where many classes of MTDs exist and they all have equal probability of success. Thus, the succeeded MTDs could have any value of O and these values represent their achieved access delay (i.e different delay values have the same probability of occurrence). As the algorithm progresses, the values of O for all MTDs increase and a shift in the delay axis (horizontal axis in Figure 4.9) happens (e.g. operating at region C). This means that the

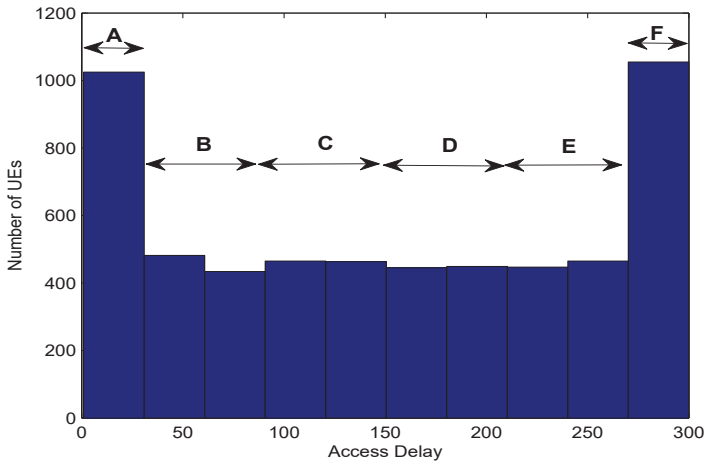

Figure 4.9: Histogram of the access delay for $N = 20000$, and $O_m = 300$

succeeded MTDs are selected from a population of MTDs with higher O values. But, again all MTDs have the same probability of success. Thus, the algorithm continues with the exact phenomena, and so on. Only, the two edges of the delay axis (i.e. regions A, F) have larger values as the congestion mainly occurs at the middle of the process. In contrast, at the start the MTDs are being served faster, and at the end the most delayed MTDs start to be served as congestion is alleviated.

To understand the impact of finite O_m on the total service time (TST), Figure 4.10 plots the TST with different O_m while increasing N. First, the figure asserts that k_e obtained from Algorithm 3 is almost the same as the value from the simulation. In addition, the figure illustrates that all TST curves saturate with the high increase of N. This is because the finite O_m value imposes an upper limit for the total service time which is equal to $I_A + O_m - 1$. This limit is the worst-case delay that could be encountered when an MTD is activated at the last activation slot I_A. However, as depicted from the results in Figure 4.10 that this upper-bound was not attained. This is usually due to the Beta-distributed traffic that results in zero activations at the very last time. Also, even if such an activation occurs, it will be served (success or block) before reaching this upper-bound since the congestion is gradually relieved as time progresses.

4.6.3 M2M-OSA with TR

For the M2M-OSA version with TR, an additional "sleep" state is considered for the devices with power consumption $P_s = 0$ mW. Using $N = 15000$, Figure 4.11 shows that the estimated N_k is very close to its actual value. One can notice that at the 100^{th} time slot the activation

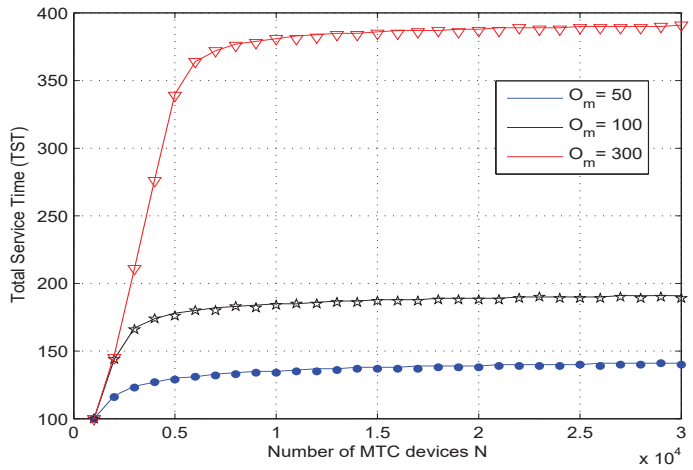

Figure 4.10: Average total service time k_e analysis vs. simulations for different number of UEs N and different O_m values. (lines: analysis, markers: sim)

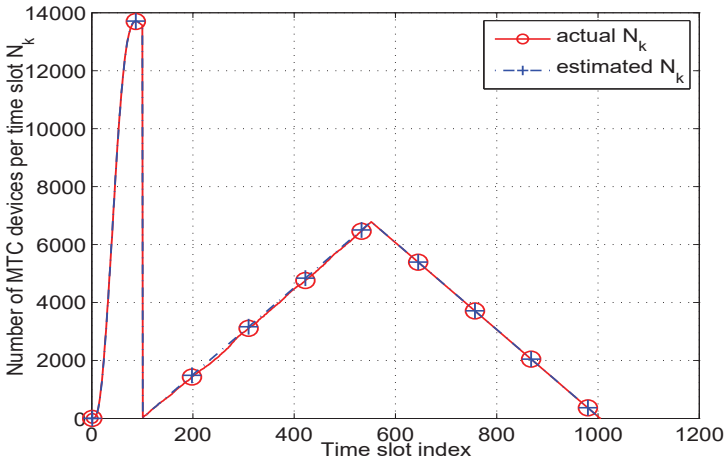

Figure 4.11: Number of M2M devices per time slot, actual vs. estimate, with traffic regulator using $f = 2$, $N = 15000$

period ends and the TR mechanism starts. So that, the value of N_k drops to zero and starts to increase again as the second activation round begins. The curves end at approximately the 1000^{th} time slot which represents the TST of the system.

Figure 4.12 first plots the average TST for the M2M-OSA in three cases, without TR, with TR and $f = 1$, and TR and $f = 2$. The figure shows that the first and third cases give almost the same performance, whereas the case of TR with $f = 1$ introduces slightly more TST. This can be interpreted by examining the definition of the traffic factor f. This factor determines the average number of devices (newly activated) at each RB. Hence, by setting $f = 1$, we set only a single device per RB in average. In reality, we may have smaller value than the average.

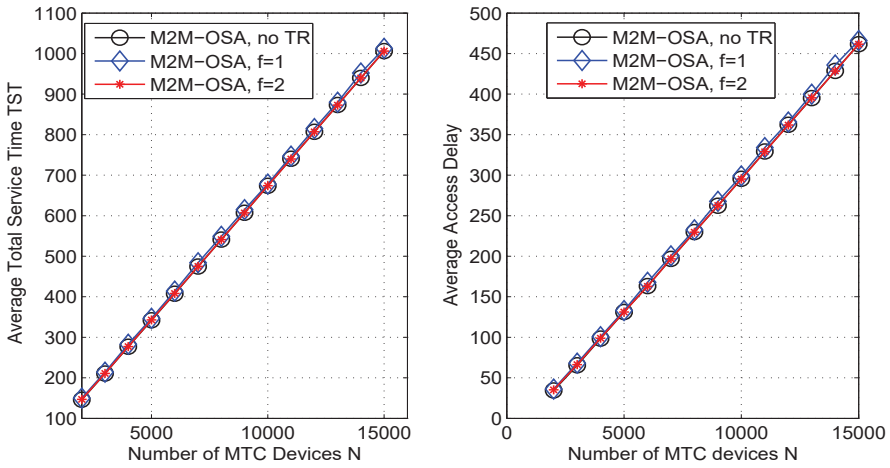

Figure 4.12: Average total service time and access delay for M2M devices while changing N for M2M-OSA with and without traffic regulation

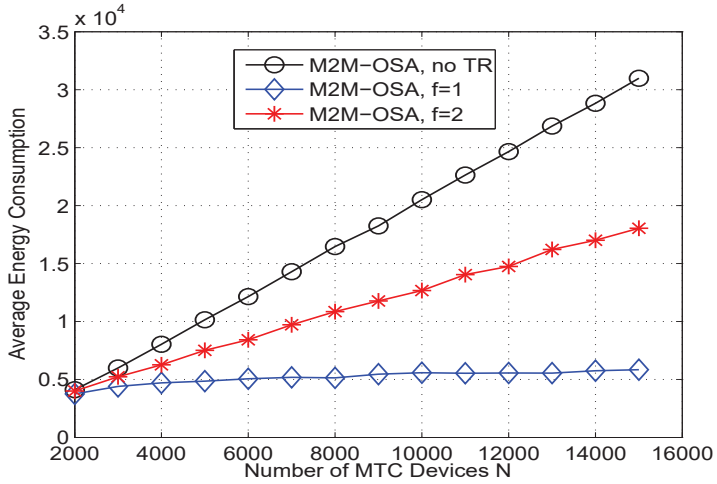

Figure 4.13: Energy consumption for M2M-OSA with and without traffic regulation

Thus, there will be some RBs without any contention which will be lost. However, this situation does not last long because more devices will enter the backlog state and the average number of devices per RB will increase. Thus, the full utilization of resources is attained until approaching similar performance as the case of no TR. Since the average TST with $f = 1$ is not much worse than the other cases, we can deduce that no much resources were lost. The figure also plots the average access delay which follows the same behavior as the TST curves.

From an energy perspective, Figure 4.13 depicts the energy consumption by each device until gaining success while changing N. Particularly, it shows the effectiveness of including the TR mechanism in saving the devices' energy and that the amount saved increases with the reduction of f. This is because decreasing f means more sleeping slots for the devices.

4.6.4 Use Cases

To sum up, M2M-OSA was shown to be effective in handling MTC with massive access. Hence, it shall be beneficial for several MTC use cases. One application is smart metering where large number of devices with small-sized data are deployed in the network. These devices autonomously send usage reports and alarm information to the smart grid infrastructure. This large volume of simultaneous transmission of delay-sensitive data could cause RAN congestion. The smart grid is one of the highest driving forces for M2M communications [123]. It provides smart capabilities to the electric grid to adapt and optimize the power generation, distribution and consumption. For such an application, long delays may lead to packet loss which could result in under- and over-supply situations (i.e. when the actual demand is different from the estimated one). These situations lead to power wastage and increased costs on the consumers [123]. According to [3], RAN congestion occurs also in a case of bridge monitoring with a mass of sensors that are triggered almost concurrently to transmit the monitoring data when a train passes on the bridge. The same thing also happens for the case of hydrology monitoring during the time of heavy rain.

Generally, event-driven services such as earthquake detection, flood detection, and disaster management services may lead to RAN overload. This is due to the mass concurrent transmissions from enormous number of devices in certain geographical area to report the incident to the remote server [124]. In [6], the 3GPP performed an evaluation to the conventional RA technique in handling the earthquake monitoring application where it was shown that these conventional mechanisms are inefficient due to the increased collisions. This in turns increases the delays which is unacceptable in the earthquake scenario, because even tens of milliseconds are very important for an earthquake alarm. Another example is fleet management where vehicles (e.g. taxis) are equipped with devices that can report their latest location information periodically or on demand. According to [6], when a large number of vehicles using fleet management service gather in a cell, overload may happen since the location information report is always in the frequency of seconds. This was noticed to be a typical scenario at certain locations such as the taxi area of the airport. Being able to achieve full resources utilization and resolve collisions, our proposed M2M-OSA is a promising candidate in handling such applications in a timely manner.

It is worth mentioning that one essential feature of MTC devices is having diverse quality of service (QoS) requirements which poses a major challenge for M2M communication to operate

effectively. Particularly, the low-power feature desired for MTC devices to save the energy limits the wireless communication performance. Moreover, the huge number of devices in M2M networks complicates the resource allocation process to be able to ensure QoS and reliability requirements of different devices. On the other hand, the mechanisms employed for M2M communication should be simple, yet effective to be capable of attaining the QoS requirements whereas at the same time preserve the low-complexity requirement for MTC devices. Efficient mode selection methods that allow the MTC devices to select either to communicate directly with the BS or via a gateway help in guaranteeing QoS [32]. Additionally, grouping based on QoS requirements can alleviate the congestion in the PRACH and hence achieve lower delays. A QoS-aware random access transmission was proposed in [33]. In [125], a resource allocation technique for M2M devices in LTE-A networks was introduced. It adopts a QoS-based cluster formation approach, then the RBs access is first granted for high-prioritized cluster with higher QoS requirement. We believe that a compromise between the complexity and the efficiency would be tailored according to the M2M application. A future direction is to include the QoS of different devices while employing M2M-OSA. Also, the algorithm can be extended by changing the competing metric to achieve certain desired criterion that is application-dependent instead of randomly scheduling any device. The inclusion of alarm traffic with higher priority is also a possible extension along with the relaxation of the ideal channel conditions assumption.

4.7 Summary

With the deployment of machine-to-machine (M2M) communications, it is expected that the number of devices will enormously increase. When these devices attempt to concurrently access the network, a radio access network overload problem arises. In this case, the conventional random access procedure used in Long Term Evolution-Advanced (LTE-A) networks is rendered inefficient due to the frequent collisions that lead to excessive delay and resource wastage. In this chapter, we formally assessed the performance of the proposed distributed overload control algorithm, namely M2M-OSA. Particularly, mathematical expressions were provided for different evaluation metrics including average access delay, total service time, energy consumption and blocking probability. To prove its effectiveness, the performance of the proposed algorithm was compared to the dynamic-access class-barring scheme where the simulation results depict the superiority of the proposed scheme. Furthermore, a binary integer

programming problem was formulated where we showed that the achieved access delay using the proposed algorithm approaches the optimal value. Moreover, we introduced a traffic regulation mechanism that achieved significant energy saving for the MTDs which is essential for battery-constrained M2M applications. We finalized by a discussion of the potential use cases of the proposed algorithms as well as a number of future directions.

5. Fast-Grant Learning-Based Approach for Machine-Type Communications with NOMA

5.1 Introduction

Thanks to the Internet of things (IoT), the communications of everyday objects become possible. In this context, machine-to-machine (M2M) communication is the baseline technology for IoT communications. According to the fifth generation (5G) wireless networks, both massive machine type communications (mMTC) and ultra reliable low latency communications (URLLC) are to be supported to enable massive and mission-critical IoT applications, respectively [126]. These two categories of M2M communications are known to have completely different features than that of the legacy human-to-human (H2H) communications. Generally, machine-type communication (MTC) is characterized by sporadic small-sized data packets with heterogeneous quality-of-service (QoS) requirements, and mostly uplink traffic. Unlike H2H, although essential, MTC is not data rate-centric networks. Instead, MTC requirements are directed to connectivity, latency, and reliability. Particularly, mMTC are about enormous number of connected devices with relaxed latency constraints. In contrast, URLLC have very strict latency and reliability requirements.

As mentioned earlier, uplink access is a fundamental challenge for MTC especially with massive connectivity. Without efficient control, when a huge number of MTC devices (MTDs) attempt to access the network simultaneously, congestion occurs. Consequently, the network becomes vulnerable to severe access delays, packet loss and even service interruption. The conventional grant-based random access (RA) used in LTE/LTE-A networks was shown to be inefficient for massive scenarios [7]. Particularly, the four-step handshaking procedure of the

RA is lengthy with expensive signaling for MTDs with small payloads. On the other hand, fully uncoordinated access, known as grant-free where the devices transmit directly on a random uplink resource without sending scheduling requests, reduces the signaling cost at the expense of collisions [15]. Having much larger number of devices than the available resources is the likely case of MTC which reflects that collisions will degrade the performance.

In the previous chapters, we proposed a distributed overload control algorithm that avoids the conventional RA scheme employed in cellular networks. We also focused on the event-driven situations where a large number of devices activate within short period of time. In general, an MTC environment is expected to have multiple types of traffic where periodic and event-driven traffic types are among the essential types [15]. Adding to this the different QoS requirements imposed by both mMTC and URLLC applications which further complicates the uplink access problem. Consequently, there is an urgent demand for the development of an effective uplink access techniques that can efficiently tackle the diverse challenges of MTC with high compatibility to the existing cellular networks. In [43], the 3GPP proposed fast uplink grant technique for latency reduction. In fast uplink grant, the base station (BS) allocates the uplink grants directly to the MTDs without receiving any scheduling requests. Hence, both large signaling overhead and collisions are avoided. Consequently, energy-constrained MTDs can significantly save energy. However, if an inactive MTD received an uplink grant, the resource is wasted.

Fast uplink grant has two main challenges [15]. First, the BS has to predict the set of active MTDs at each cycle. Second, the granted resources should go to the optimal MTDs according to each network requirements. This should be done with limited or no information of the devices' QoS requirements or channel state information (CSI) which are very difficult to acquire for massive number of MTDs at the BS. Learning techniques are known to be very promising in solving problems in such environments. In [127], a traffic prediction framework is provided for MTDs activated by binary Markovian events where the resources are allocated using fast grant to MTDs that have the maximal likelihood for transmission. The results depicted the superiority of fast grant as compared to the conventional RA. However, the results incorporated relatively small number of MTDs while ignoring their QoS during the scheduling step. In [128], multi-armed bandit (MAB) learning technique was adopted to solve the scheduling problem. In spite of the promising results, the performance of the technique in [128] is dependent on the prediction accuracy. Hence, there is a need to decouple the two challenges of fast grant schemes

or even reduce the effect of the predictor efficiency on the scheduler performance.

In recent years, non-orthogonal multiple access (NOMA) has received great attention as a promising enabling technique for beyond 5G wireless networks [16], [12]. As compared to orthogonal multiple access (OMA) where every user is exclusively allocated the resource, NOMA allows multiple users to share he same spectrum resource at the same time while multiplexing them either in code or power domain. This permits serving more users using limited resources in less time. For MTC, NOMA is an excellent candidate for improving the spectral efficiency, reducing latency and efficiently handling massive connectivity. These promising gains come at the expense of a more complex receiver that is able to decode the superposed signals. For uplink NOMA, successive interference cancellation (SIC) decoder is used at the BS which is acceptable for low-budget MTDs with mostly-uplink traffic.

In this chapter, we propose NOMA-based fast uplink grant for MTC to enhance the overall system performance and approach the decoupling of the predictor and scheduler performances. By allowing multiple MTDs to share the same resource, the resource wastage due to prediction errors can decrease. Obviously, this comes with the cost of extra signaling and complexity. However, starting with the simple 2-user NOMA can efficiently handle the tradeoff between system performance and complexity. For proper SIC decoding, certain level of distinctness between the received signals at the BS should be maintained [68]. This was usually done by the BS in a centralized manner such that the users' channels are sorted and distinctive ones are paired together (e.g. [16]). Clearly, for massive networks, sending this immense amount of CSI information to the BS is impractical. Hence, we provide a two-step NOMA-based fast uplink grant scheme that employs MAB learning for fast scheduling of the MTDs. Afterwards, distinctive NOMA pairing is done in a distributed manner such that SIC can successfully operate. To facilitate the problem investigation, we first provide a source traffic prediction abstraction model. This abstraction model eases incorporating the predictor performance in the network to be able to study its effect on the fast uplink scheduler and assess the overall network performance. Finally, we introduce a BIP formulation for the optimal NOMA pairing to illustrate the further enhancements that could be attained from NOMA-based fast uplink grant.

5.2 Related Works

Due to the advantages of NOMA in improving the spectral efficiency, handling massive connectivity and reducing latency, several works in literature studied NOMA with IoT and 5G [11], [16]. In [126], a power-domain uplink NOMA access scheme was proposed to solve a joint sub-carrier and power allocation problem for narrow-band IoT (NB-IoT). The goal is to maximize the number of connected MTDs under QoS and power constraints. In [129], a code-domain NOMA scheme was proposed to support massive MTC. In this scheme, each MTD randomly selects a sub-channel to transmit its data. Then, each device selects a channel code with appropriate code rate to encode its message along with its terminal identity (ID). Since the number of MTDs selecting the same sub-channel is not fixed, Raptor codes, which are rateless codes, were used. In this case, collisions only occur when multiple MTDs select the same sub-band and code structure.

In literature, variants of NOMA with MTC could be found. In [130], a random access-based NOMA scheme for M2M networks was proposed. By modifying the conventional 4-step random access (RA) scheme used in LTE, this scheme allows multiple MTDs to share the same resource block (RB). Power control was also employed so that the BS could not only detect the presence of the preambles from the timing advance information but also estimate the number of MTDs that have chosen each preamble. Then, the MTDs apply code-domain NOMA to transmit their data over the same channel. In contrast, other works apply NOMA with grant-free scenarios to avoid the large overhead of the RA process [14]. Grant free is employed to allow the users to transmit directly whenever they have data without waiting for a grant from the BS. In [7], the authors analyzed the performance of a grant-free NOMA MTC system in terms of throughput and outage probabilities assuming both successive joint decoding (SJD) and SIC under a Rayleigh fading and path loss channel model. In the adopted scheme, multiple MTDs transmit their data associated with a signature pilot sequence, selected from a set of orthogonal pilots, in the same channel. In [61], a distributed layered grant-free NOMA framework was proposed to mitigate the inherited collisions of grant-free systems. In this case, the cell is divided into layers with different power levels, and only devices of different layers could share the same RB without collisions. To further reduce the collisions, enhanced access barring (EAB) was applied.

Although grant-free avoids the lengthy handshaking process of the conventional RA, there

is a complete lack of control resulting in higher probability of collisions with MTC massive access. In [131], a semi grant-free with NOMA proposal was introduced where a channel is granted for a high priority MTD via conventional grant-based RA. Then, grant-free MTDs are permitted to share the channel provided that the performance of the grant-based MTD is not too much degraded. As indicated before, although variants of NOMA with MTC ranging from grant-based to grant-free are introduced in the literature, there is almost no work that combined NOMA with fast uplink grant. To the best of our knowledge this is among the very first work to study fast uplink grant for MTC with NOMA. Moreover, centralized user pairing is usually applied in NOMA scenarios which may not be suitable for massive access [68, 135]. Hence, in this paper, we propose a heuristic distributed pairing scheme.

5.3 System Model

We consider a single cell network where N MTDs are served by a single BS. Assume that the available bandwidth is divided into M orthogonal RBs, each of width B. Assuming slotted time, at each time slot (i.e. cycle) the BS allocate the RBs to the available "active" MTDs using fast grant. The term "active" refers to an MTD with a packet ready for transmission (i.e. non-empty buffer) within a predefined maximum tolerable delay identified by the IoT application. If a packet waits for access more than its maximum tolerable delay, it will be dropped out. A packet's access delay is defined as the number of cycles elapsed from the moment at which the packet is ready for transmission until it is granted an uplink resource. Due to the heterogeneity of IoT applications, the packets have different QoS requirements that are unknown to the BS.

To model congestion, at the beginning, we consider a Beta-distributed activation of the MTDs as suggested by the 3GPP for overload situations [6]. Particularly, the N MTDs are activated within a bounded activation time T_A that is assumed to be divided into I_A time slots. Considering a Beta distribution with parameters $(\alpha = 3, \beta = 4)$, each MTD is activated at time t with probability $A(t)$ as follows:

$$A(t) = \frac{t^{\alpha-1}(T_A - t)^{\beta-1}}{T_A^{\alpha+\beta-1}\mathcal{B}(\alpha, \beta)}, \tag{5.1}$$

where $\mathcal{B}(\alpha, \beta)$ is the Beta function [114]. This more realistic model is only adopted during the first activation of the devices instead of the uniform activation model adopted in [128]. Then, at each cycle, a random set of inactive MTDs is selected for activation to join the pool of

backlogged MTDs.

All channel gains are assumed to be independent and identically distributed (i.i.d) across users and time. Also, the link gain is assumed to be fixed during the whole cycle. Particularly, let $h_l = a_l g_l$ is the channel gain between the l^{th} MTD and the BS [62]. $g_l \sim \mathcal{CN}(0,1)$, $a_l = 10^{\frac{a_{l,dB}}{10}}$ represent the small scale Rayleigh fading, and the large scale fading, respectively. With path loss PL_{dB} and log-normal shadowing X_σ with variance σ, the large scale fading is represented as $a_{l,dB} = PL_{dB} + X_\sigma$. Finally, The 3GPP path loss model of $PL_{dB} = 128.1 + 37.6\log(d_{BS})$ is adopted where d_{BS} is the distance to the BS [132].

For uplink NOMA network, multiple MTDs transmit non-orthogonally to the same receiver (i.e. BS). The received signals at the BS are then decoded using SIC. Hence, certain level of *distinction* has to be maintained. For 2-user NOMA, the received signal at the BS on the m^{th} RB is as follows:

$$\mathbf{y}_m = \sqrt{P_s} h_{sm} \mathbf{x}_{sm} + \sqrt{P_w} h_{wm} \mathbf{x}_{wm} + \mathbf{z}_m \tag{5.2}$$

where P_s, h_{sm}, P_w, h_{wm} are the transmission power and channel gain for the strong and weak MTDs of the m^{th} NOMA pair, corresponding to their signals $\mathbf{x}_{sm}, \mathbf{x}_{wm}$, respectively. $\mathbf{z}$ is an additive white Gaussian noise with zero mean and power spectral density N_0. We also assume that each MTD is able to acquire the CSI of its link to the BS from the pilot signal sent from the BS and it has a maximum power budget P_t.

Using SIC, the BS decodes the strong signal first, then removes it and decodes the weak signal. Hence, only the strong MTD will suffer from interference. Based on that, the achievable rates of the NOMA pair can be expressed as:

$$r_s = B\log_2\left(1 + \frac{P_s\gamma_s}{P_w\gamma_w + 1}\right) \tag{5.3}$$

$$r_w = B\log_2\left(1 + P_w\gamma_w\right) \tag{5.4}$$

where $\gamma = \frac{|h|^2}{N_0 B}$ is the normalized channel gain. The subscripts $(.)_s, (.)_w$ denote the strong and weak MTDs, respectively.

5.4 Problem Formulation

For massive number of MTDs, the resource allocation process is known to be complicated due to the limited resources. Additionally, the diverse applications of IoT increases the process

complexity due to the heterogeneous QoS requirements of the MTDs. In such a network, the objective is to schedule the set of MTDs that maximizes the achieved utility under QoS constraints. Based on that, the optimal OMA scheduler can be formulated as:

$$\begin{aligned} \mathcal{S}(t) = \underset{\{i_1,\ldots,i_M\}\in\mathcal{K}(t)}{\arg\max} \quad & \textstyle\sum_{i=i_1}^{i_M} U_i(t) \\ \text{s.t.} \quad & r_i(t) \geq R_{i_{min}}, \quad i = i_1,\ldots,i_M \\ & d_i(t) \leq D_i, \quad i = i_1,\ldots,i_M \end{aligned} \tag{5.5}$$

where $\mathcal{K}, \mathcal{S}$ are the sets of active and scheduled MTDs, respectively. $U_i(t)$ is the utility of the scheduled MTD i. $R_{i_{min}}, D_i$ are the minimum rate and maximum delay constraints of MTD i, whereas $r_i(t), d_i(t)$ are the achieved rate and access delay for MTD i at t. For massive MTC, delay requirements are commonly the most important. However, other QoS metrics can be imposed. Hence, the utility function U is defined as a combination of different normalized QoS metrics as [128]:

$$U_i(t) = \delta_1 v_i(t) + \delta_2 r_i^n(t) + \delta_3 f(D_i(t)), \tag{5.6}$$

where $\delta_1, \delta_2, \delta_3$ are weights of each metric summing up to 1. At time t, $v_i(t)$ is the value of information of the generated packet by MTD i at certain context [133]. The normalized value represents a percentage of importance, hence $v_i(t) \in [0, 1]$. Also, $r_i^n(t)$ is the normalized rate of MTD i that is obtained by dividing the achieved rate $r_i(t)$ by the maximum rate R_{max} that could be achieved by the node having the best channel to the BS. In our setting, an arbitrary relatively-high value for R_{max} was selected. For the maximum tolerable access delay $D_i(t)$, the normalization was obtained using a modified Gompertz function with parameters a, b, c as follows:

$$f(D_i(t)) = a - ae^{-be^{-cD_i(t)}}. \tag{5.7}$$

Figure 5.1 shows the modified Gompertz function with different values of the parameters a, b, c. It shall be noted that the lower the maximum tolerable delay, the higher the function value which reflects higher reward by selecting MTDs with more strict delay requirements as desired.

Solving the problem in (5.5) requires the BS to gather information about the MTDs in the network. For instance, the BS needs to know the QoS requirements of each active MTD at each time instant, as well as the CSI for optimal throughput. However, gathering such information for massive number of devices, usually with small-sized data packets, is highly inefficient. In such

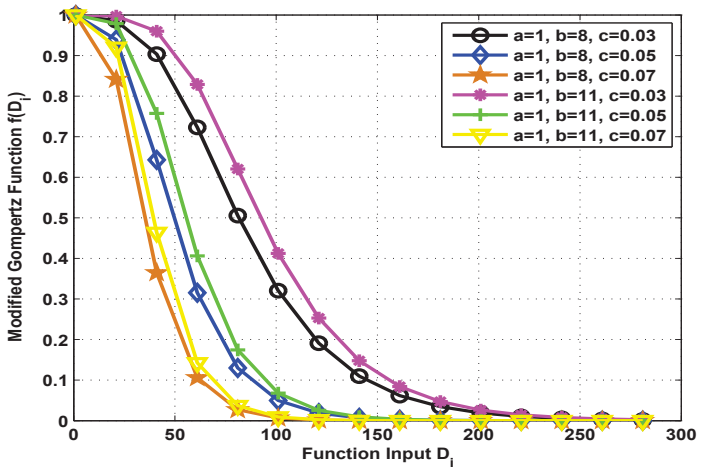

Figure 5.1: Modified Gompertz function for modeling latency with different values of the control parameters

uncertain environments, learning techniques are good candidates to solve the problems with limited or no information at the decision maker. In the following, online learning is employed to solve the problem along with NOMA to boost the performance.

5.5 Proposed Two-stage NOMA-based Fast Grant Scheme

To efficiently apply fast uplink grant in MTC networks, the associated challenges related to both scheduling and activity prediction should be tackled. The proposed scheme aims to tackle this by applying MAB learning to partially schedule the MTDs and learn the utilities over time with no information at the BS related to MTDs' channels or QoS [128]. Then, NOMA technique is adopted to enhance the overall system performance, reduce latency, and to better employ the limited resources. The proposed scheme comprises of two phases as follows.

(1) MAB is used such that the BS could select the scheduled MTDs for fast uplink grant.

(2) The scheduled MTDs are considered as cluster heads (CHs) and seek pairing with other "nearby" devices that satisfy a NOMA power tolerance condition announced by the BS for successful SIC operation.

After transmission, the BS receives QoS-based rewards from the scheduled MTDs. In this scheme, we allow inactive CHs that received an uplink grant, while having no ready-for-transmission packets, to seek pairing as well. This implies that the CHs can forward the grant to other "active" MTDs that respond to the pairing request. This procedure assists in avoiding the re-

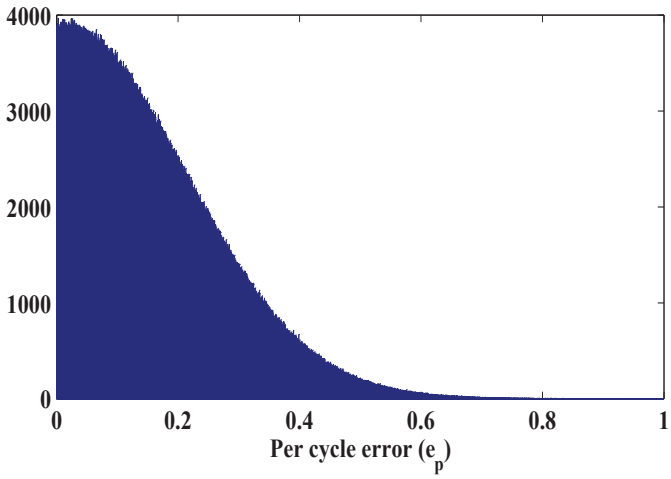

Figure 5.2: Truncated histogram of the distribution of the prediction error $e_p \sim \mathcal{N}(0.01, 0.04)$

source wastage inherited in fast uplink grant. In the following, we show how the preceding traffic prediction step is abstracted. Then, detailed description of each phase is provided.

5.5.1 Traffic Predictor Abstraction

We assume that the BS employs a predictor with certain average prediction error $\bar{e}_p$. At each cycle, there are an actual active MTDs list and a BS prediction list that satisfies a per-cycle error $e_p \sim \mathcal{N}(\bar{e}_p, \sigma_e^2)$ truncated in [0,1]. We define e_p as the total number of errors in prediction to the total number of MTDs in the network. Thus, for a network of N MTDs, we have:

$$e_p = \frac{N_e}{N} = \frac{N_{a,i} + N_{i,a}}{N} \tag{5.8}$$

where N_e is the total number of prediction errors per cycle. Also, $N_{a,i}, N_{i,a}$ are the number of active MTDs that were predicted as inactive, and the number of inactive MTDs predicted as active by the BS, respectively. Furthermore, we assume that the predictor outputs a probability of being active P_a for each MTD. Figure 5.2 illustrates the histogram of the distribution of $e_p \sim \mathcal{N}(0.01, 0.04)$.

It shall be noted that since the adopted error model is truncated normal distribution in $[0, 1]$, the actual mean of e_p, is calculated as follows:

$$\tilde{e}_p = \bar{e}_p + \sigma_e \frac{\phi(a_e) - \phi(b_e)}{\Phi(b_e) - \Phi(a_e)} \tag{5.9}$$

where $a_e = -\bar{e}_p/\sigma_e, b_e = (1 - \bar{e}_p)/\sigma_e$. Also, $\phi(\cdot), \Phi(\cdot)$ are the pdf of the standard normal

distribution and its cumulative distribution function, respectively. Using $\tilde{e}_p$, the expected value of prediction errors can be obtained as follows:

$$E\{N_e\} = N\tilde{e}_p. \tag{5.10}$$

5.5.2 Fast Grant Scheduling using MAB

MAB is a classic decision making problem in which a set of predefined arms exists where a decision maker selects an arm and observes the reward. Although the rewards are drawn from some unknown distributions for the decision maker, the goal is to maximize the cumulative system reward. Due to lack of information, the selected arm may be an inferior arm in terms of the average reward. Hence, another performance measure for MAB problems is the *regret* which is defined as the difference between rewards of the best arm that could have been played and the selected arm. In our problem, the BS is the decision maker and the MTDs are the arms. According to (5.5), the reward of the i^{th} MTD is defined as:

$$\theta_i(t) = \mathbb{1}[r_i(t) \geq R_{i_{min}}]\mathbb{1}[d_i(t) \leq D_i]U_i(t), \tag{5.11}$$

where $\mathbb{1}(.)$ is an indicator function equal 1 if its argument holds and 0 otherwise. It is used to assure that both minimum rate $R_{i_{min}}$ and maximum tolerable delay D_i requirements of the i^{th} MTD are satisfied. In terms of regret, the goal of the scheduler is to minimize the cumulative regret $\mathcal{R}$. Let $\theta_k(t)$ be the achieved reward of playing arm k at time t, and $\theta^*(t)$ be the maximum reward that could have been achieved at time t, the regret up to time T is then given by:

$$\mathcal{R}(T) = \mathbb{E}\left[\sum_{t=1}^{T}\theta^*(t) - \sum_{t=1}^{T}\sum_{k\in\mathcal{S}(t)}\theta_k(t)\right]. \tag{5.12}$$

To maximize the cumulative reward, the upper-confidence bound (UCB) concept is used to solve the problem. This method is well-known to achieve balance between exploitation and exploration tradeoff [134]. This well-known machine learning tradeoff requires balancing reward maximization based on exploiting the knowledge already acquired while attempting (i.e. exploring) new actions to further enhance knowledge. Since the availability of the MTDs is probabilistic, sleeping MAB is more suitable [128]. In sleeping MAB, at each time instant, only a subset of the arms is available. Given that the BS employs a prediction algorithm, at

each time slot t, it has the predicted set of active MTDs $\mathcal{K}'_t$ associated with certain probability of being active $P_{ai}(t)$ for each MTD i in $\mathcal{K}'_t$. Then, the BS employs the UCB to play an arm $k(t)$ such that:

$$k(t) = \arg\max_{i \in \mathcal{K}'_t} P_{ai}(t) \left(\frac{z_i(t)}{n_i} + \sqrt{\frac{8 \log t'}{n_i}} \right), \tag{5.13}$$

where $z_i(t)$ is the sum of rewards of MTD i up to time t, n_i is the number of times MTD i was selected while being active, and t' is the total number of times the selected MTD was active. Note that all the parameters in (5.13) are known to the BS except for t'. Using the traffic predictor, the estimate of t' is assumed to be the number of times each MTD was estimated active by the BS. In this scheme, the BS selects the newly activated MTDs first before starting (5.13). Then, it starts to schedule the MTDs with the highest UCB values. Note that, the newly activated MTDs are the ones that were estimated as active at the BS, and yet they have no learning history.

5.5.3 NOMA Pairing

Although MAB learning has the potential to boost the scheduling efficiency of fast uplink grant with limited information, its performance is highly dependent on the preceding prediction step. For OMA, the spectrum resource is completely wasted in case the selected MTD by the BS was actually inactive. With NOMA, multiple MTDs are allowed to share the same spectrum resources which enhances the spectrum utilization. Furthermore, a NOMA-based network would be less susceptible to traffic prediction errors. Hence, a 2-user NOMA scheme is adopted. Note that, this scheme could be extended to allow multiple MTDs per RB with an increase in the system complexity.

The proposed NOMA scheme starts by identifying the CHs which are the MTDs granted the uplink resources from the preceding MAB stage. These CHs will seek pairing to other active MTDs in the network that were not granted an uplink resource, namely non-CHs (nCHs). In particular, the pairing is initiated by each CH through sending a pairing request containing its ID and a signal-to-noise ratio (SNR) threshold, γ_{th}. This threshold guarantees the minimum required distinction between the signals of the paired devices at the BS for proper SIC operation. Afterwards, the CHs will receive responses from their *eligible* nCHs. In this regard, we assume that each nCH associates itself to the nearest CH for a chance to be paired with it. Also, we refer to the set of nCHs associated to the same CH as cluster. If γ_{th} is satisfied, the CH will

randomly select one of its responded nCHs to share the frequency band. It is worth mentioning that the nearest CHs can be identified from the received signal strength (RSS) of their pairing requests. Although the association assumption is restrictive, it reduces the communication overhead needed for the distributed pairing by forcing a single eligible CH for each nCH which is reasonable for massive MTC. Moreover, this assumption helps to reduce the interference and allows the CHs to lower the power needed for sending pairing requests. However, it could be easily relaxed by allowing each nCH to associate itself to multiple CHs according to their RSS.

Assuming that the BS announces a tolerance power, P_{tol}, defined as the minimum power difference required for efficient SIC operation, γ_{th} can be derived from:

$$P_s\gamma_s - P_w\gamma_w \geq P_{tol}, \tag{5.14}$$

where P_s, γ_s are the transmission power and the normalized channel gain of the strong MTD of certain NOMA pair and the BS, respectively. Similarly, P_w, γ_w are for the weak MTD of the same NOMA pair. Assuming equal power allocation where $P_s = P_w = P_t$, then γ_{th} is:

$$\gamma_{th} = \begin{cases} \gamma_{CH} + \frac{P_{tol}}{P_t}, & \text{if } \gamma_{CH} = \gamma_w(weak), \\ \gamma_{CH} - \frac{P_{tol}}{P_t}, & \text{if } \gamma_{CH} = \gamma_s(strong). \end{cases} \tag{5.15}$$

Hence, an MTD n can be paired to its nearest CH c iff:

$$\gamma_n \geq \gamma_{th}, \quad \text{if } \gamma_{th} = \gamma_c + \tau, \;\; mode = 0, \tag{5.16}$$

$$\gamma_n \leq \gamma_{th}, \quad \text{if } \gamma_{th} = \gamma_c - \tau, \;\; mode = 1. \tag{5.17}$$

where $\tau = \frac{P_{tol}}{P_t}$. An extra control bit $mode$ could be added to identify what type of members the CH is seeking. For instance, $mode\ 0$ can be used for throughput maximization such that the CHs seek pairing to stronger MTDs. However, being decoded at last stage, the CH is susceptible to SIC error propagation unless high level of distinction is maintained among the pairs to reduce outages. For a network of both mMTC and ultra reliable low latency communications (URLLC) MTDs, $mode = 1$ is recommended to protect the CH selected by the BS that is most probably would be URLLC MTD with service priority. In general, if a network operates on certain mode but some CHs did not find eligible pairs, it may be allowed to switch mode to find a pair.

Under the prescribed scenario, if an inactive CH was granted a resource, it will seek pairing

as well. Thus, the resource is not wasted with an incentive of an increase in the CH's cumulative reward. Particularly, with NOMA, the total system reward at each cycle is the sum resulted from both CHs and nCHs which could reach double the achieved rewards with OMA. However, the following cumulative rewards of each CH c and its nCH pair n are used in (5.13):

$$z_c(t) = z_c(t-1) + \theta_c(t) + \rho\theta_n(t) \tag{5.18}$$

$$z_n(t) = z_n(t-1) + (1-\rho)\theta_n(t), \tag{5.19}$$

where ρ is a weight factor defining the share of the nCH's reward belonging to the paired CH as an incentive. For a general MTD i, the values of other parameters in (5.13) are adjusted independently based on its individual reward θ_i. For instance, n_i is only incremented for non-zero reward MTDs. This setting overcomes the prediction inaccuracy as the active nCHs themselves respond to the pairing request. Figure 5.3 illustrates the procedure of the proposed two-stage scheme.

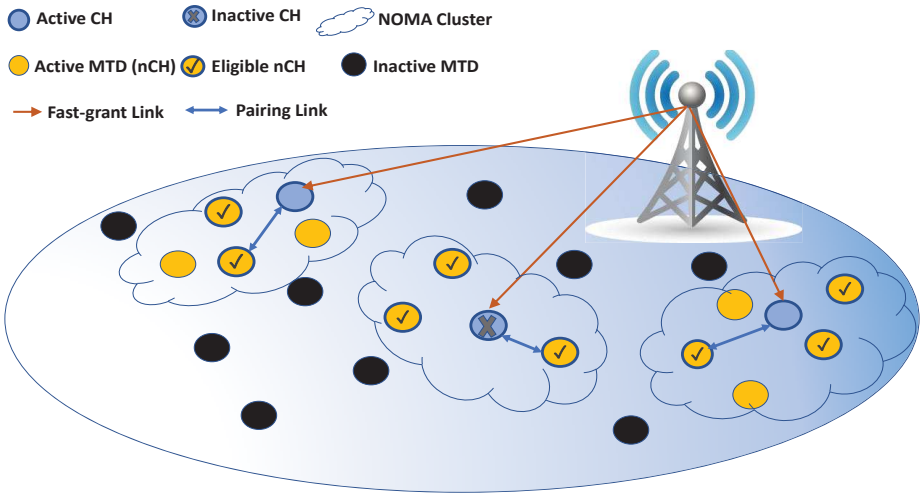

Figure 5.3: An illustration of the proposed two-stage NOMA-based fast grant scheme with 3 RBs whereas the paired MTDs are going to share the same RB using NOMA

5.6 Optimal NOMA Pairing

The regret defined in (5.12) gives an indication of how far the system performance is from the optimal performance. For OMA system, the optimal scheduling is to schedule the MTDs with the highest rewards at each cycle. In this case, the BS should have all the information about all the devices to schedule the set that maximizes the total network reward. This information

includes all devices' CSI, and QoS requirements which is impractical for massive number of devices.

For our proposed NOMA scenario, in order to get a meaningful regret, we need to find the optimal NOMA pairing that maximizes the network reward to be the reference performance. This optimal pairing is known to be complex for massive MTC [135]. Particularly, for M RBs, the number of pairing sets grows super-exponential with M, renders the optimal solution infeasible even for moderate M. However, our objective is to find a reference performance that is comparable to the proposed two-stage scheme to specifically help in improving the second stage related to NOMA pairing. Specifically, the first stage related to MAB was adopted since it helps to improve the scheduling quality with limited or no information at the BS. Then, NOMA was used to further improve the resources utilization. Given that we keen to keep the benefits of MAB while at the same time improve the NOMA-pairing stage, we formalize a *quasi-optimal* NOMA scenario. In this scenario, the CHs selected at the first step would be inputs to the problem and it is required to find the optimal pairs for the given CHs. This formulation helps to assess the performance of the proposed NOMA with random pairing so that we can improve our proposed 2-step scheme in the future.

Based on the above, in obtaining optimal pairing, we allow the BS at each cycle to grant the M resources either to the best active CHs whose rewards are the highest or to M CHs selected with MAB. Then, these selected CHs enter a second stage to find their optimal pair. The problem is formalized as a binary integer programming (BIP) problem with the objective of maximizing the total system reward:

$$\max \quad \sum_{c=1}^{M} \sum_{n=1}^{N_{an}} \omega_{c,n} I_{c,n} \tag{5.20}$$

$$\text{s.t.} \quad \sum_{n=1}^{N_{an}} I_{c,n} \leq 1, \quad \forall c \tag{5.21}$$

$$\sum_{c=1}^{M} I_{c,n} \leq 1, \quad \forall n \tag{5.22}$$

$$I_{c,n} \in \{0,1\}, \quad \forall n, c \tag{5.23}$$

$$|P_n \gamma_n - P_c \gamma_c| \geq P_{tol}, \quad \forall n, c \tag{5.24}$$

where $I_{c,n}$ is a pairing binary variable that takes the value 1 when CH c is paired with nCH n. The nCH n index is set to span all actual active N_{an} nCHs, whereas a maximum of M CHs, already selected at the first step, exist in the network at each cycle. The first two conditions are to assure that each CH is paired to only one nCH and vice versa. The last condition is the pairing

condition required for successful SIC decoding, and the absolute value is used to indicate both modes $0, 1$. The objective of the optimization problem is to maximize the reward resulted from pairing. This is done with the optimization weights $\omega_{c,n} = \theta_c + \theta_n$. To satisfy the last condition, we could set $\omega = 0$ while solving the problem if the condition does not hold for any pair. The problem was solved using Matlab.

5.7 Simulations

In the following, we consider N MTDs randomly located at fixed points of a square area with side length 500 meters. At the beginning, the activation of the MTDs follows the Beta distribution with parameters ($\alpha = 3, \beta = 4$) within $I_A = 10$ slots. After the first activation of the N MTDs, we randomly select MTDs for reactivation at each cycle to keep a dynamic activation process. Note that by considering a case of $N >> M$, the overload situation sustains. For the channel modeling parameters, a noise power is considered to be -174 dBm/Hz, bandwidth is 360 kHz and standard deviation for the log normal shadow fading is 10 dB [128]. The traffic predictor is assumed to have $e_p \sim \mathcal{N}(0.01, 0.04)$, and $P_a \in [0.8, 1]$.

Regarding the utility function (5.6), we set $\delta_1 = 0.2, \delta_2 = 0.3, \delta_3 = 0.5$, where the delay gets the highest weight. Additionally, we set the parameters of the function in (5.7) as $a = 1, b = 8, c = 0.03$ [128]. To better illustrate the ability of MAB learning in achieving the essential delay requirement of MTC, we chose to set the first $N/2$ MTDs to be with strict maximum delay of $D_i \in [1, 100]$ slots, whereas the remaining $N/2$ MTDs are with relaxed maximum delay of $D_i \in [150, 300]$ slots. Running the simulations for a total of T cycles, the maximum delay of the MTDs changes with every activation. Also, we assumed that the rate threshold is satisfied for all MTDs (i.e. $\mathbb{1}[r_i(t) \geq R_{i_{min}}] = 1$ in (5.11)). Furthermore, all the results are generated assuming mode 0.

Finally, the system performance is evaluated in terms of the rewards and regret defined in (5.11), (5.12). Additionally, we defined a measure for resource wastage namely the missing ratio as the ratio between the number of missed resources to the total number of available resources during the whole period (i.e. MT). The access delays and the number of scheduled devices will also be illustrated. Only non-zero reward selected MTDs are counted in calculating different system metrics. Table 5.1 shows the parameters used in the simulations.

Figure 5.4 plots the histogram of the number of times each MTD was scheduled for both

Table 5.1: System model parameters

Parameter	Definition	Value
M	Number of uplink resources (RBs)	10
N	Number of MTDs	500
T	Total number of cycles	10000
P_t	Transmission power	10 dBm
P_{tol}	Detection threshold for SIC	4 dBm
ρ	Weight for nCH reward division	0.3

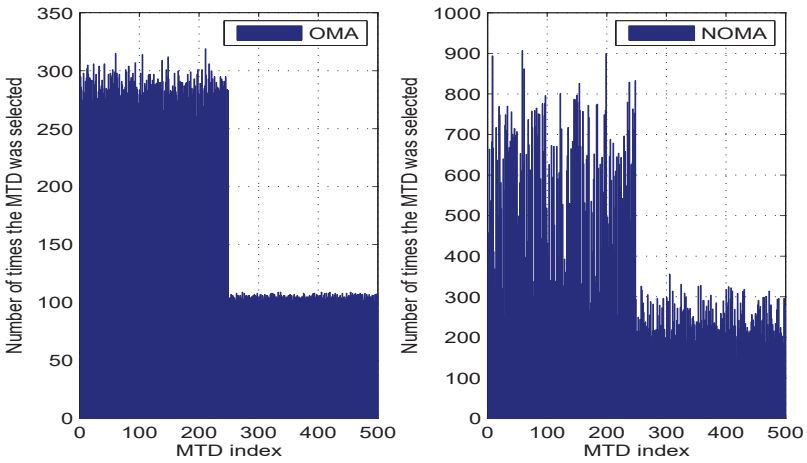

Figure 5.4: Number of times each MTD was scheduled in both OMA and NOMA systems

OMA and NOMA systems. In both cases, it is shown that the MAB learning technique is able to schedule the MTDs with strict delay requirements more frequently (i.e. first $N/2$ MTDs). However, with NOMA, the total number of scheduling times is higher. This is due to the better utilization of the resources offered by NOMA. Figure 5.5 shows the division of the scheduled devices in NOMA case. Particularly, it shows the number of times each MTD was scheduled as CH or nCH. As the figure depicts, the increase in the number of scheduled CHs reflects the improvement achieved in the MAB performance. At the same time, the CHs' scheduling follows the same MAB's trend where MTDs with strict delays are scheduled more frequently. On the other hand, the nCH MTDs are selected by each CH randomly from the eligible active devices. Hence, the figure shows that all N MTDs are scheduled uniformly. At the end, the overall scheduling performance is enhanced to offer better utilization of the resources. This was also verified in Figure 5.6 where the accumulated number of missed resources with time for both cases are plotted. The figure depicts the enhancement achieved by the proposed NOMA in utilizing the resources missed due to prediction errors. The staircase appearance of NOMA case is a result of the infrequent wastage of the resources.

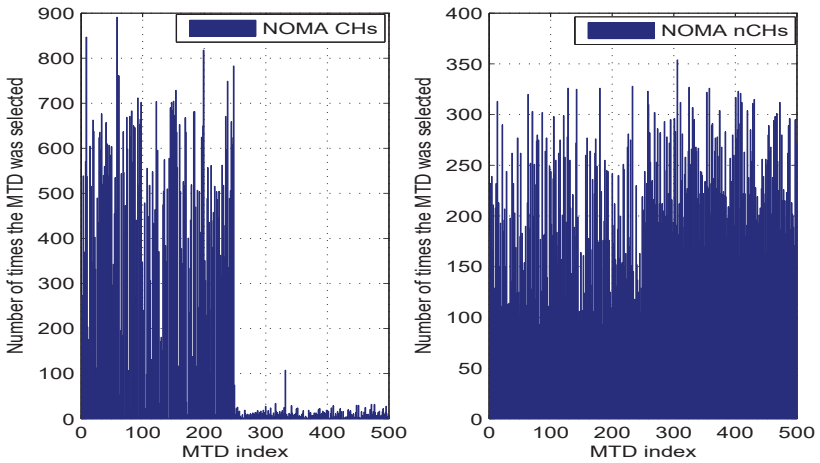

Figure 5.5: Number of times each MTD was scheduled in NOMA scenario as CH (left) or nCH (right)

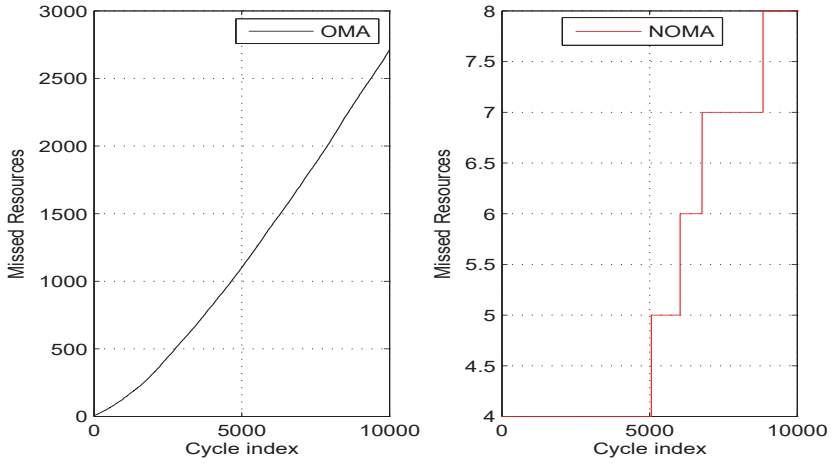

Figure 5.6: Total waste of resources in both OMA and NOMA systems

Regarding the rewards and regret, in Figure 5.7, we take the optimal rewards achieved by scheduling the available highest-reward MTDs, one per each RB, as the minuend of (5.12). This is the curve labeled as "Best" in Figure 5.7-(a). The subtrahends are the achieved rewards with MAB only and MAB with NOMA for OMA and NOMA regrets, respectively. Figure 5.7-(a) depicts the improvement gained from NOMA that exceeds the best OMA performance. This can also be seen from the regret shown in Figure 5.7-(b). Although the reference optimal reward here is not related to NOMA, it is good to see how far our practical proposed NOMA-MAB scheme is from the impractical best OMA case that needs the gathering of all MTDs' information and QoS requirements at the BS to be able to optimally allocate the resources.

In terms of the missing ratio, Table 5.2 compares the performance of OMA and NOMA.

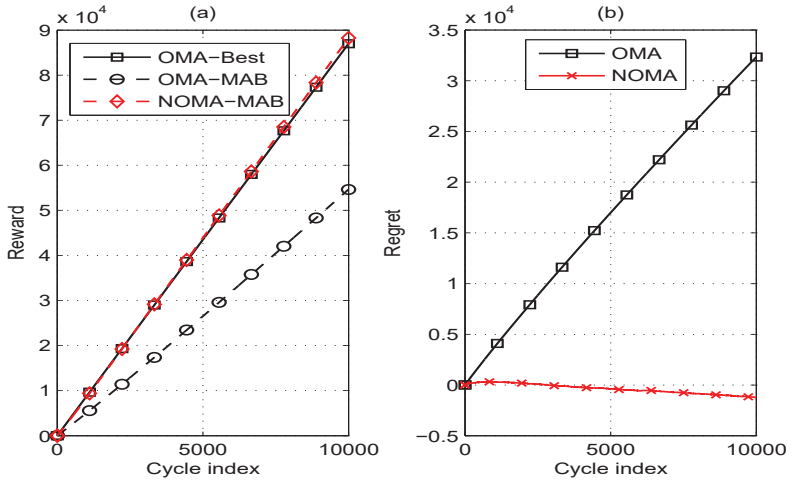

Figure 5.7: (a) Rewards resulted from both OMA and NOMA systems where best reward is calculated with at most 1 MTD/RB. (b) The difference (Best-MAB) gives the regret.

Table 5.2: Performance Evaluation with t' estimated

System	Missing Ratio	Winners	Avg. Max Delay	Avg. Access Delay
OMA	0.0271	9.7288×10^4	105.1120	36.6666
NOMA	0.0001	19.0483×10^4	108.5102	19.0421
NOMA-CHs		9.1158×10^4	59.0349	10.2293

The table depicts that NOMA offers better utilization of the resources where the missing ratio significantly decreases. It also shows the achieved increase in the number of served MTDs (i.e. winners) along with an enhancement in their average access delay. The uniform selection of nCHs pairs results in a slight increase in the average maximum tolerable delay of the scheduled MTDs. The last row shows the performance of CHs ONLY with NOMA. The results reflect an improvement in the MAB performance itself that enables faster scheduling of MTDs (less access delay) with more strict delay requirements. Note that, the modification proposed in (5.18), (5.19) allows faster learning by scheduling more MTDs at each cycle. Particularly, the proposed NOMA reduces the time wasted in scheduling the newly activated devices first in order to build their history by storing the nCH UCB parameters in (5.13) as well as the CHs'.

Focusing on NOMA, Figure 5.8-(a) compares the reward resulted from the proposed NOMA with simple random pairing to the rewards resulted from the 2-step optimal NOMA (i.e. quasi optimal) pairing analyzed in section 5.6. We consider two inputs for optimal pairing, the first is the best CHs with the highest rewards, and the second is the MAB-selected CHs. These input CHs will seek pairing using (5.20). The gaps between each of these two curves and the proposed NOMA represent the further enhancement that can be achieved by combining NOMA

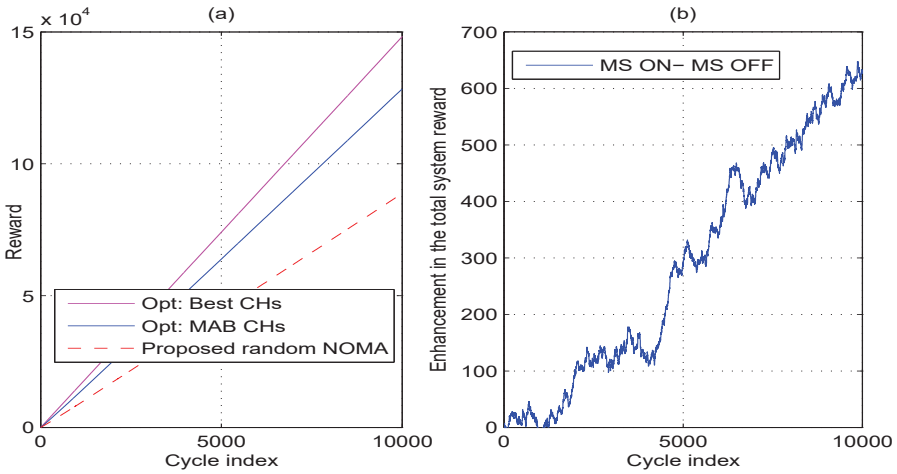

Figure 5.8: Further enhancements to NOMA system (a) Comparison of proposed and optimal NOMA scenarios (b) Effect of mode switch (MS) function on the system reward

with fast grant in a 2-stage fashion as proposed in this chapter. Specifically, the gap between the proposed NOMA curve and the optimal with MAB CHs indicates the further enhancement that could be achieved by enhancing the pairing scheme. On the other hand, to enhance the performance of random NOMA, we added a mode switch (MS) function to the system. This function allows each CH to switch its pairing mode temporarily if it failed to find a NOMA pair with the nominal mode of the network. Specifically, for a network operating at $mode = 0$, each CH that does not find a pair at certain cycle is allowed to re-announce itself as a CH operating at $mode = 1$ at this specific cycle. This is to increase the resources utilization and not to miss the pairing opportunity. Figure 5.8-(b) depicts the difference in the system reward while the function is ON and OFF. However, this enhancement results in an increase of the pairing overhead and time. The mode switch function could be useful for networks with strict pairing conditions or non-dense devices where the probability of not finding a pair increases.

Using different average prediction errors $\bar{e}_p = 0.01, 0.1, 0.4$, the effect of the predictor efficiency for both OMA and NOMA is examined. Figure 5.9 shows a relatively large increase in the number of wasted resources of OMA system with the increase of $\bar{e}_p$ which is not the case for NOMA. This reflects that NOMA system is less vulnerable to the predictor efficiency. For the rewards illustrated in Figure 5.10, although NOMA reward is always higher than OMA with different $\bar{e}_p$, increasing $\bar{e}_p$ degrades the performance of both systems. For OMA, the performance degradation is due to the loss of the resources. In contrast, with NOMA the reason is the random selection of the nCHs which represent larger portion of the selected MTDs with high $\bar{e}_p$. This can be improved by enhancing the employed pairing scheme. Overall, the results show

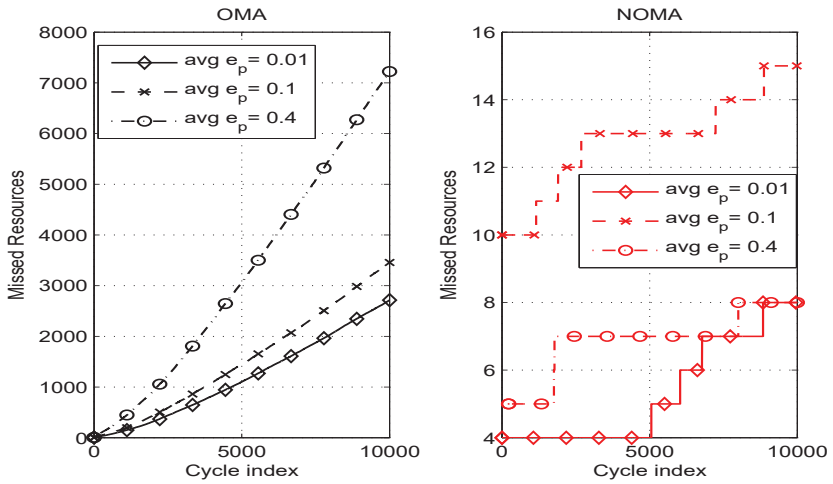

Figure 5.9: Effect of the prediction error on the resource wastage in both OMA and NOMA systems

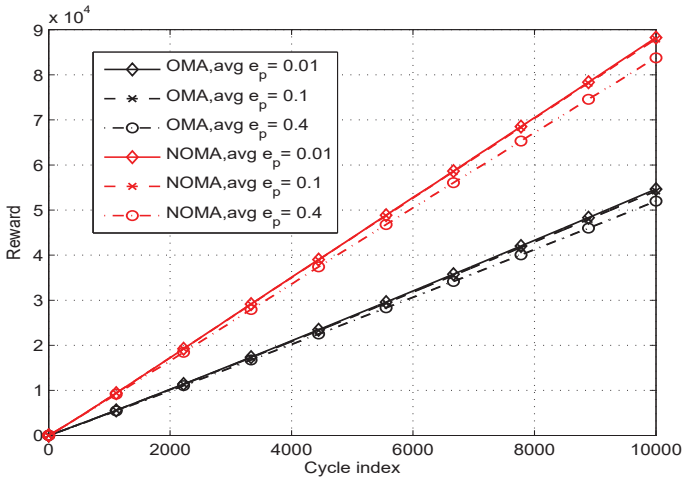

Figure 5.10: Effect of the prediction error on the system reward in both OMA and NOMA systems

a potential of performance improvement at an affordable complexity.

5.8 Summary

MTC are known for its challenging radio access network (RAN) design due to massive connectivity. In this chapter, we proposed a NOMA-based communication framework that allows MTDs to access the network while avoiding congestion. The proposed technique is a two-stage mechanism that first employs fast uplink grant to schedule the devices without sending a request to the BS. Secondly, NOMA pairing is employed in a distributed manner to reduce the signaling overhead. Due to the limited capability of information gathering at the BS in massive scenarios, learning techniques are best fit for such problems. Thus, MAB learning was adopted

to schedule the fast grant MTDs taking into account their heterogeneous QoS requirements including latency. Then, constrained random NOMA pairing was proposed to assist in decoupling the two main challenges of fast uplink grant schemes, namely active set prediction and optimal scheduling. Using NOMA, we were able to significantly reduce the resource wastage due to prediction errors. Additionally, the simulation results show that the proposed scheme can easily attain the impractical optimal OMA performance, in terms of the achievable rewards, at an affordable complexity. It was also depicted that the proposed scheme is more robust against source traffic prediction errors. A quasi-optimal 2-stage NOMA formulation was provided as a benchmark for future performance enhancement.

6. Analysis of Resources Wastage in Fast Uplink Grant for MTC with NOMA

6.1 Introduction

Efficient uplink access is a keystone for successful deployment of MTC that enables the promising Internet of things (IoT). The conventional random access (RA) technique and its variants were shown to suffer from excessive collisions and prohibitive overhead such that MTC latency and reliability requirements cannot be guaranteed. In the previous chapter, we adopted fast uplink grant access technique for MTC which was first proposed by the 3GPP for latency reduction [43]. In fast uplink grant, the MTDs receive uplink grants allocated directly from the BS without the need to transmit scheduling requests. This technique avoids both collisions and signaling overhead. Consequently, energy-constrained MTDs can significantly save energy. However, there is a risk of wasting the valuable resources in case the grant is allocated to an inactive MTD. Mainly, there are two challenges associated to fast uplink grant [15]. First, the prediction of the set of active MTDs at each cycle. Second, the optimal MTDs scheduling such that the resources are allocated to the best MTDs according to each network's criteria. Due to the difficulty of information gathering for enormous number of MTDs, scheduling should be done with limited or no information of the devices' QoS and channel state information (CSI) at the BS.

In chapter 5, to comply with the aforementioned requirements, we proposed NOMA-based fast uplink grant scheme that employs MAB learning for scheduling a first incumbent of the spectrum resource, referred to as CH. Afterwards, each CH seek pairing to other "active" MTDs to share the scheduled resources in a NOMA fashion under certain constraints. Particularly, a minimum power tolerance should be maintained among the paired MTDs to guarantee distinction for successful SIC operation. The preliminary results of this technique showed promising

enhancements in terms of rewards, access delays, and resource utilization.

In this chapter, we theoretically assess the performance of the proposed NOMA-based fast grant uplink scheme. To simplify the analysis, minor modifications were adopted in the system model. First, we ignore the large-scale fading parameters and considered only small scale Rayleigh fading. Particularly, all the channel gains are modeled as Rayleigh fading distributed as $\mathcal{CN}(0,1)$, and assumed to be independent and identically distributed (i.i.d) across users and time. In addition, after the I_A time slots of the activation period, we assume that exactly N_a randomly-selected MTDs are active at each cycle. Mostly important, we apply a modification to the prediction error model that eases the exact calculation of the probability of selecting inactive MTDs. Based in that we provide a complete analytical framework for the probability of resource wastage of both OMA and NOMA systems taking into consideration the decoding order of the SIC receiver. The outage probabilities of the MTDs are also derived under different scenarios while accounting for the activity of the scheduled MTDs. Moreover, we compare the proposed scheme to its OMA and random-scheduling NOMA counterparts and show its superiority in terms of rewards, delay, and resource wastage.

6.2 Prediction Error Model

Fast uplink grant requires the BS to know the set of active MTDs at each cycle such that the scheduled MTDs can be selected. We assume that the BS employs a source traffic predictor for this task. Furthermore, we assume that the predictor outputs a probability of being active P_a for each MTD along with the set of active MTDs. We abstract the source traffic prediction stage by characterizing its prediction performance in terms of missed-detection, and false-alarm probabilities. Particularly, the probability of missed detection is defined as the probability that an active MTD is predicted as inactive. Also, the probability of false alarm is defined as the probability that an inactive MTD is declared as active. Based on that, we assume that the BS employs a predictor that achieves a per-device missed-detection, P_n^m, and false-alarm, P_n^f, such that their corresponding system-level probabilities are $P_M = \sum_{n=1}^{N} P_n^m/N$ and $P_F = \sum_{n=1}^{N} P_n^f/N$, respectively [52]. Given that N_a MTDs are active at each cycle, this modeling reflects a per-cycle average number of missed-detection, and false-alarm events equal $\mu_{a,i} = N_a P_M$, $\mu_{i,a} = (N - N_a)P_F = N_i P_F$, respectively [52]. Hence, following this model, we arbitrary assume that a per-cycle missed-detection and false-alarm errors can be generated using

Table 6.1: Division of MTDs

Prediction \ Actual	**Active**	**Inactive**	**Sum**
Active	$N_{a,a}$	$N_{i,a}$	N_{BS}
Inactive	$N_{a,i}$	$N_{i,i}$	
Sum	N_a	N_i	N

truncated normal distributions as follows:

$$N_{a,i} \sim \mathcal{N}(\mu_{a,i}, N_a^2\sigma^2) \qquad \in [0, N_a] \tag{6.1}$$

$$N_{i,a} \sim \mathcal{N}(\mu_{i,a}, N_i^2\sigma^2) \qquad \in [0, N_i] \tag{6.2}$$

where $N_{a,i}, N_{i,a}$ are the number of active MTDs that were predicted as inactive, and the number of inactive MTDs predicted as active by the BS, respectively. Also, σ is an arbitrary variable to adjust the distribution variance. Based on this error model, Table 6.1 presents the division of the MTDs based on their actual and predicted status where the shaded entries represent errors. Note that, the purpose of this abstraction modeling is to have, at each cycle, an actual active MTDs list and a BS prediction list that satisfies certain prediction error performance for simulation purposes. A real system would have an actual predictor that generates the BS prediction list based on certain real information.

As mentioned earlier, with fast uplink grant the resource is wasted if it is allocated to an inactive MTD. Hence, we proceed to find the probability of selecting an inactive MTD, p_i. Let M_e be the number of RBs allocated to inactive MTDs given that M RBs are available, hence the expected value $E\{M_e\}$ is obtained as:

$$\begin{aligned} E\{M_e\} &= \sum_{m=1}^{M} m \times \mathbb{P}(M_e = m) \\ &= \sum_{m=1}^{M} m \times \frac{\binom{N_{i,a}}{m}\binom{N_{a,a}}{M-m}}{\binom{N_{BS}}{M}}. \end{aligned} \tag{6.3}$$

Since, $N_{i,a}$ and $N_{a,i}$ are truncated normal random variables, we use the law of total probability

to calculate $\mathbb{P}(M_e = m)$, and to find p_i as follows:

$$\mathbb{P}(M_e = m) = \sum_{n_{a,i}=1}^{N_a} \sum_{n_{i,a}=1}^{N_i} \frac{\binom{n_{i,a}}{m}\binom{n_{a,a}}{M-m}}{\binom{n_{BS}}{M}} \times \mathbb{P}(N_{a,i} = n_{a,i})\mathbb{P}(N_{i,a} = n_{i,a}) \tag{6.4}$$

$$p_i = \frac{E\{M_e\}}{M} \tag{6.5}$$

where $n_{a,a}, n_{BS}$ can be easily found using Table 6.1. Note also that the combination conditions must be satisfied.

6.3 Performance Analysis of the Proposed Scheme

In this section, we analyze the performance of the proposed two-phase scheme in terms of the average number of wasted resources. In this regard, a RB is considered to be wasted if its overall achieved reward is zero. Hence, to obtain the probability of wasting a resource using fast uplink grant, the following probabilities are needed:

$$\begin{aligned} p(\theta_i = 0) &= p(r_i < R_{i_{min}}) + p(d_i > D_i) + p(U_i = 0) \\ &= P_{outage} + P_{inactiveMTD} + P_{U0} \end{aligned} \tag{6.6}$$

From (5.6), it can be assumed that $P_{U0} \neq 0$ for any arbitrary MTD. Thus, in what follows, we focus on evaluating both the outage and inactive-scheduled probabilities. It is worth mentioning that we assume an MTD is inactive if it has no data for transmission which could be a result of exceeding its packet maximum tolerable delay. Moreover, we assume that all MTDs have the same R_{min}, and a normalized noise power $N_0 B = 1$.

6.3.1 OMA-Based Fast Grant

In OMA, as indicated by (6.6), the RB is wasted if it is allocated to an inactive MTD or if the active scheduled MTD is in outage. This can be formulated for the selected MTD c as follows:

$$\mathbb{P}(\theta_c = 0) = \mathbb{P}(r_c < R_{min} | c\ active)\mathbb{P}(c\ active) + \mathbb{P}(c\ inactive), \tag{6.7}$$

where $\mathbb{P}(c\ inactive)$ is evaluated as in (6.5). Also, the outage probability can be evaluated as follows:

$$\begin{aligned}\mathbb{P}(r_c < R_{min}) &= \mathbb{P}(\log(1+P_t|h_c|^2) < R_{min}) \\ &= \mathbb{P}(|h_c|^2 < \frac{\varepsilon}{P_t}) \\ &= 1 - e^{-\varepsilon/P_t}, \end{aligned} \tag{6.8}$$

where $\varepsilon = 2^{R_{min}} - 1$. The expression in (6.8) is the CDF of the exponential random variable $|h_c|^2 \sim \exp(1)$. Finally, the probability of wasting a resource in OMA, P_w^{OMA}, is given by:

$$P_w^{OMA} = (1-p_i)(1-e^{-\varepsilon/P_t}) + p_i \tag{6.9}$$

6.3.2 NOMA-Based Fast Grant

For NOMA networks, to obtain the probability of wasting a resource, P_w^{NOMA}, we differentiate between two cases depending on whether the selected CH is active or not. In this regard, P_w^{NOMA} can be expressed as follows:

$$P_w^{NOMA} = \mathbb{P}(\text{waste}|\text{active CH})\mathbb{P}(\text{active CH}) + \mathbb{P}(\text{waste}|\text{inactive CH})\mathbb{P}(\text{inactive CH}), \tag{6.10}$$

where $\mathbb{P}$(inactive CH), $\mathbb{P}$(active CH) are $p_i, 1-p_i$ as in (6.5), respectively. With inactive CH granted the fast uplink resource, the resource is wasted iff the CH does not find a pair or if the found pair is in outage. Note that an inactive CH does not restrict pairing based on channel conditions since the CH will not transmit anyways. Consequently, both modes, 0 and 1, give the same resource wastage result in case of inactive CHs. With random NOMA pairing, any nCH in the cluster is eligible to be granted the resource. Based on that, the probability of not finding a pair, P_{NP}, in this case is the probability of having an empty cluster as follows:

$$P_{NP}|\text{inactive CH} = \mathbb{P}(N_{c_i} = 0) = (1 - \frac{1}{M})^{N_{ar}}, \tag{6.11}$$

where N_{c_i} is the number of MTDs in cluster C_i such that $N_{c_i} \sim \text{Bin}(N_{ar}, 1/M)$ for nearest association to the CHs. Also, N_{ar} is the number of active nCHs. Particularly, using (6.3), one can obtain the expected number of remaining active MTDs (i.e. nCHs) per cycle after fast grant

selection of the CHs, $E\{N_{ar}\}$ as:

$$E\{N_{ar}\} = N_a - (M - E\{M_e\}). \tag{6.12}$$

The second scenario is to find a pair but an outage occurs to the paired nCH. This probability can be expressed for nCH, n, as follows:

$$\begin{aligned} P_o^{nCH}|\text{inactive CH} =& \mathbb{P}(r_n < R_{min}|\text{inactive CH}) \\ =& \mathbb{P}(\log(1 + P_t|h_n|^2) < R_{min}, N_{c_i} \neq 0) \\ =& \left(1 - e^{-\varepsilon_t}\right)\left(1 - (1 - \frac{1}{M})^{N_{ar}}\right), \end{aligned} \tag{6.13}$$

where $\varepsilon_t = (2^{R_{min}} - 1)/P_t$. Finally, the probability of wasting a resource given inactive CH is evaluated as:

$$P_w^{NOMA}|\text{inactive CH} = P_{NP}|\text{inactive CH} + P_o^{nCH}|\text{inactive CH}. \tag{6.14}$$

Next, we study the cases of wasting the resources with active CH scheduled for fast grant which depend on the operating mode. It shall be emphasized that in the proposed pairing approach, at first each nCH associate itself to the nearest CH and then it checks its eligibility to be paired with the associated CH. The pairing condition is as follows:

$$P_s\gamma_s - P_w\gamma_w \geq P_{tol}. \tag{6.15}$$

where P_s, γ_s are the transmission power and the normalized channel gain between the strong MTD of certain NOMA pair and the BS, respectively. Similarly, P_w, γ_w are for the weak MTD of the same NOMA pair. All the channel gains are modeled as Rayleigh fading that are i.i.d across users and time, $|h_s|^2, |h_w|^2$ are exponential random values with unit rate. Assuming equal power allocation where $P_s = P_w = P_t$, the pairing condition can be rewritten as:

$$|h_s|^2 - |h_w|^2 > \tau \tag{6.16}$$

where $\tau = \frac{P_{tol}}{P_t}$. Depending on the decoding order of both the CH and nCH, we defined the two operating modes 0,1. In particular, in mode 0, the CH is the weak MTD that is decoded last. Also, in mode 1, the CH is the strong MTD that is decoded first. Based on that, MTD n can be

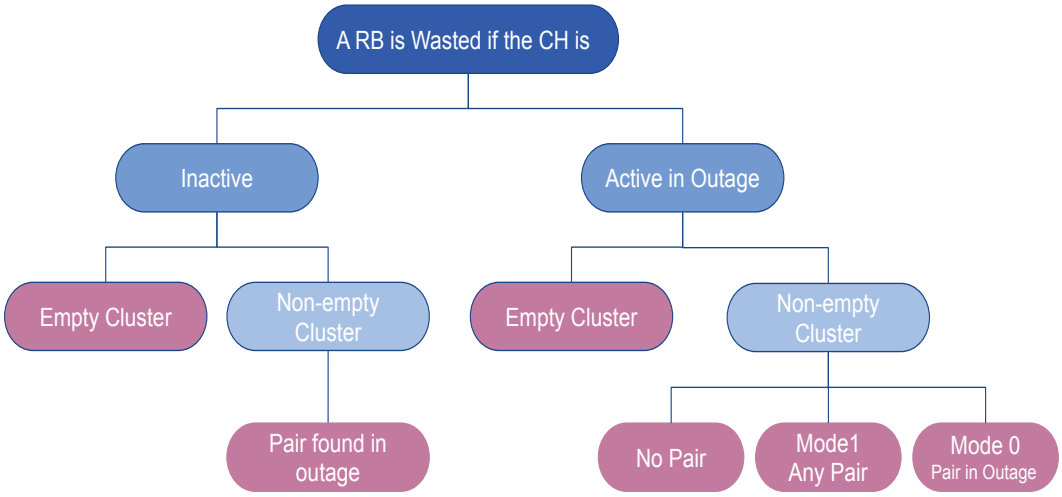

Figure 6.1: Probability tree diagram of the resources wastage from the proposed NOMA scheme

paired to its nearest CH c iff:

$$|h_n|^2 \geq |h_c|^2 + \tau, \qquad mode = 0 \tag{6.17}$$

$$|h_n|^2 \leq |h_c|^2 - \tau, \qquad mode = 1. \tag{6.18}$$

Figure 6.1 illustrates the probability tree diagram of the wasted resources resulted from the proposed NOMA scheme. Although it may seem that the two modes of operation are almost identical except at a single case, the pairing conditions make the analysis (i.e the outage calculations) of both modes completely different as will be shown in the next subsections.

6.3.2.1 NOMA-Based Fast Grant: Mode 1

In mode 1, the CH is the strong MTD that is decoded first and seek pairing to other weaker MTDs. Hence, if the CH is in outage the resource is wasted whether a pair is found or not (i.e. $P_w^{NOMA}|\text{active CH} = P_o^{CH}$). Specifically, even if a nCH pair is found, the nCH signal cannot be correctly decoded due to SIC error propagation. However, we study each case separately since the CH's achievable rates and conditions are different. We begin by the following expression for an active CH outage:

$$P_o^{CH} = P_o^{CH}{}_{|N_{c_i}=0}\mathbb{P}(N_{c_i} = 0) + P_o^{CH}{}_{|N_{c_i}\neq 0}\mathbb{P}(N_{c_i} \neq 0), \tag{6.19}$$

where the outage of the first term with empty cluster (i.e. $P_o^{CH}{}_{|N_{c_i}=0}$) is the same as OMA in (6.8). Regarding the term of nonempty cluster, we divide the outage into two cases whether a pair is found or not. Hence, $P_o^{CH}{}_{|N_{c_i}\neq 0} = P_{o,P}^{CH} + P_{o,NP}^{CH}$ where the non-empty cluster notation was ignored for notation simplicity. Starting by the case of not finding an eligible pair (i.e. $P_{o,NP}^{CH}$), this means that among N_{c_i} members of the cluster C_i, none satisfied the pairing condition in (6.15). Since the paired nCH n is the weak MTD, $|h_n|^2 \leq |h_c|^2 - \tau$ must hold as indicated in (6.18). Assuming the channel gains of the nCHs are sorted as follows $|h_1|^2 < |h_2|^2 < \cdots < |h_{N_{c_i}}|^2$, $P_{o,NP}^{CH,1}$ can be formulated as:

$$\begin{aligned} P_{o,NP}^{CH,1} =& \mathbb{P}\left(\log(1+P_t|h_c|^2) < R_{min}, |h_1|^2 > |h_c|^2 - \tau\right) \\ =& \mathbb{P}\left(|h_c|^2 < \varepsilon_t, |h_c|^2 < |h_1|^2 + \tau\right), \end{aligned} \tag{6.20}$$

where the superscript $(.)^1$ indicates mode 1. Given that $|h_1|^2 \sim \exp(N_{c_i})$ and is independent of $|h_c|^2$, the joint probability density function (PDF) $f_{|h_1|^2,|h_c|^2}(h_1, h_c)$ is given by:

$$f_{|h_1|^2,|h_c|^2}(h_1, h_c) = N_{c_i} e^{-N_{c_i} h_1} e^{-h_c}, \quad h_1, h_c > 0 \tag{6.21}$$

Using (6.21), $P_{o,NP}^{CH,1}$ can be obtained as:

$$P_{o,NP}^{CH,1} = \begin{cases} 1 - e^{-\varepsilon_t}, & \text{for } \tau \geq \varepsilon_t, \\ 1 - \dfrac{N_{c_i} e^{-\tau}}{1+N_{c_i}} \left(1 - e^{-(1+N_{c_i})(\varepsilon_t - \tau)}\right) - e^{-\varepsilon_t} e^{-N_{c_i}(\varepsilon_t - \tau)}, & \text{for } \tau < \varepsilon_t. \end{cases} \tag{6.22}$$

The next studied case is the outage of active CH with nonempty cluster where a pair is found. The CH's outage probability $P_{o,P}^{CH,1}$ is calculated as follows:

$$\begin{aligned} P_{o,P}^{CH,1} =& \mathbb{P}\left(\log(1 + \frac{P_t|h_c|^2}{P_t|h_n|^2+1}) < R_{min}, |h_c|^2 - |h_n|^2 > \tau\right) \\ =& \mathbb{P}\left(|h_c|^2 < \varepsilon(|h_n|^2 + \frac{1}{P_t}), |h_c|^2 > |h_n|^2 + \tau\right). \end{aligned} \tag{6.23}$$

Although all channel gains are assumed to be i.i.d with $\exp(1)$, the employed pairing condition affects the channel gain distribution of the selected nCHs where both $|h_c|^2$ and $|h_n|^2$ are correlated [131]. Therefore, conditioned on $|h_c|^2$, all eligible nCHs have the following cumulative

distribution function (CDF):

$$F^1_{|h_n|^2}(h_n) = \begin{cases} \frac{1-e^{-h_n}}{1-e^{\tau-h_c}}, & \text{for } h_n < h_c - \tau \\ 1, & \text{for } h_n > h_c - \tau. \end{cases} \tag{6.24}$$

The corresponding PDF for the eligible nCHs, denoted by $f^1_{|h_n|^2}(h_n)$, can be obtained as:

$$f^1_{|h_n|^2}(h_n) = \frac{e^{-h_n}}{1-e^{\tau-h_c}}, \qquad \text{for } h_n < h_c - \tau. \tag{6.25}$$

Using mathematical manipulations, and considering the case of $\varepsilon > 1$, the outage probability in (6.23) can be expressed as:

$$P^{CH,1}_{o,P} = e^{-\tau}\Big[e^{-\frac{\varrho}{\varepsilon}} \int_0^{e^{\frac{\varrho}{\varepsilon}-q}} \ln|1-t^{\varepsilon}|\mathrm{d}t + (1-e^{-q})\ln|1-e^{-q}| + e^{-q}\Big],\ \varepsilon > 1, \tau > \varepsilon_t, \tag{6.26}$$

where $\varrho = \tau - \varepsilon_t$, and $q = \varrho/(\varepsilon - 1)$. For the case of $\tau < \varepsilon_t$, we take the limit of (6.26) as $q \to 0$, and $P^{CH,1}_{o,P}$ is expressed as:

$$P^{CH,1}_{o,P} = e^{-\tau}\Big[e^{-\frac{\varrho}{\varepsilon}} \int_0^{e^{\frac{\varrho}{\varepsilon}}} \ln|1-t^{\varepsilon}|\mathrm{d}t + 1\Big],\ \varepsilon \geq 1, \tau \leq \varepsilon_t. \tag{6.27}$$

Note that, for the cases of $\varepsilon < 1$, the outage probability either equal to zero or has very small value that can be neglected, specifically when τ is very close to ε_t. Now, all the factors that contribute in resource loss for fast uplink grant scenario with NOMA are obtained and the waste probability, P^{NOMA}_w, can be calculated as in (6.10) using (6.14), (6.19).

To complete the analysis, we also study the outage probability of the nCHs. Particularly, nCH outage with inactive CH pair was given in (6.13). Additionally, nCH outage while being paired with active CH can be expressed as:

$$\begin{aligned} P^{nCH,1}_{o,P} =& 1 - \mathbb{P}\Big(\log(1+P_t|h_n|^2) \geq R_{min}, \log(1+\frac{P_t|h_c|^2}{P_t|h_n|^2+1}) \geq R_{min}, |h_c|^2 - |h_n|^2 > \tau\Big) \\ =& 1 - \mathbb{P}\Big(|h_n|^2 \geq \varepsilon_t, |h_c|^2 \geq \varepsilon(|h_n|^2 + \frac{1}{P_t}), |h_c|^2 > |h_n|^2 + \tau\Big). \end{aligned} \tag{6.28}$$

As before, due to the pairing condition imposed in (6.28), the PDF of $|h_n|^2$ is as indicated in

(6.25). Using mathematical manipulation, the nCH outage with active pair is obtained as:

$$P_{o,P}^{nCH,1} = 1 + e^{-\tau-\frac{\varrho}{\varepsilon}} \int_0^{e^{\frac{\varrho}{\varepsilon}-Q}} \ln|1-t^{\varepsilon}|\mathrm{d}t - e^{-\tau}\Big(c_1 \ln|c_1| - c_2 \ln|c_2| - c_1 + c_2\Big), \tag{6.29}$$

where $\varrho = \tau - \varepsilon_t$, $Q = (q, \varepsilon_t)^+$, and $q = \varrho/(\varepsilon - 1)$. Also, $c_1 = 1 - e^{-\varepsilon_t}$, and $c_2 = 1 - e^{-Q}$.

It is also important to evaluate the pairing success probability with active CH, defined as the probability that each cluster has eligible nCH pairs (i.e. $N_{g_{c_i}} \neq 0$) among which the scheduled nCH is selected. Given that a cluster C_i has N_{c_i} MTDs with channel gains $|h_1|^2 < |h_2|^2 < \cdots < |h_{N_{c_i}}|^2$, and using the pairing condition in (6.18), we have:

$$\mathbb{P}(N_{g_{c_i}} = n_g) = \mathbb{P}\left(|h_{n_g}|^2 \leq |h_c|^2 - \tau, |h_{n_g+1}|^2 > |h_c|^2 - \tau\right). \tag{6.30}$$

Therefore, the probability of finding at least one eligible pair (i.e. $N_{g_{c_i}} \neq 0$) is calculated using the joint PDF in (6.21) as:

$$\begin{aligned}\mathbb{P}(N_{g_{c_i}} \neq 0) =& 1 - \mathbb{P}\left(|h_1|^2 > |h_c|^2 - \tau\right) \\ =& \frac{N_{c_i} e^{-\tau}}{N_{c_i}+1}.\end{aligned} \tag{6.31}$$

6.3.2.2 NOMA-Based Fast Grant: Mode 0

In mode 0, the scheduled CH is the weak MTD which is decoded after its nCH pair. Also, with all MTDs having the same R_{min} and due to SIC error propagation, the resource is wasted if the nCH pair is in outage. Therefore, in this case $P_w^{NOMA}|$active CH is expressed as:

$$P_w^{NOMA}|\text{active CH} = P_o^{CH}{}_{|N_{c_i}=0}\mathbb{P}(N_{c_i} = 0) + P_{w|N_{c_i}\neq 0}\mathbb{P}(N_{c_i} \neq 0). \tag{6.32}$$

$$P_{w|N_{c_i}\neq 0} = P_{o,NP}^{CH,0} + P_{o,P}^{nCH,0} \tag{6.33}$$

where $P_o^{CH}{}_{|N_{c_i}=0}$ is the probability of CH outage given it has an empty cluster (i.e. no nCHs) which is calculated as in (6.8). Furthermore, (6.33) represents the waste in case of nonempty cluster which results either by CH outage while having no pair, $P_{o,NP}^{CH,0}$, or nCH outage with pair, $P_{o,P}^{nCH,0}$. It shall be noted that the superscript $(.)^0$ refers to mode 0. Using the pairing condition in (6.17), and assuming the channel gains of the nCHs are sorted as follows $|h_1|^2 < |h_2|^2 <$

$\cdots < |h_{N_{c_i}}|^2$, $P_{o,NP}^{CH,0}$ can be formulated as:

$$\begin{aligned} P_{o,NP}^{CH,0} =& \mathbb{P}\Big(\log(1+P_t|h_c|^2) < R_{min}, |h_{N_{c_i}}|^2 < |h_c|^2 + \tau\Big) \\ =& \mathbb{P}\Big(|h_c|^2 < \varepsilon_t, |h_c|^2 > |h_{N_{c_i}}|^2 - \tau\Big), \end{aligned} \tag{6.34}$$

where $|h_{N_{c_i}}|^2$ is the maximum channel gain among N_{c_i} values. Hence, the following are its CDF, $F_{|h_{N_{c_i}}|^2}$, and the pdf, $f_{|h_{N_{c_i}}|^2}$:

$$\begin{aligned} F_{|h_{N_{c_i}}|^2}(h) =& \left(1-e^{-h}\right)^{N_{c_i}} \\ f_{|h_{N_{c_i}}|^2}(h) =& N_{c_i}\left(1-e^{-h}\right)^{N_{c_i}-1} e^{-h}, \quad h \geq 0. \end{aligned} \tag{6.35}$$

Given that $|h_{N_{c_i}}|^2, |h_c|^2$ are independent, their joint pdf is defined in the first quadrant as $f_{|h_{N_{c_i}}|^2}(h) \times e^{-h_c}$. Therefore, $P_{o,NP}^{CH,0}$ can be obtained as:

$$P_{o,NP}^{CH,0} = \frac{e^{\tau}}{N_{c_i}+1}\left[(1-e^{-\tau-\varepsilon_t})^{N_{c_i}+1} - (1-e^{-\tau})^{N_{c_i}+1}\right]. \tag{6.36}$$

The second term in (6.33) is $P_{o,P}^{nCH,0}$ which is given by:

$$\begin{aligned} P_{o,P}^{nCH,0} =& \mathbb{P}\left(\log(1+\frac{P_t|h_n|^2}{P_t|h_c|^2+1}) < R_{min}, |h_n|^2 - |h_c|^2 > \tau\right) \\ =& \mathbb{P}\left(|h_n|^2 < \varepsilon(|h_c|^2 + \frac{1}{P_t}), |h_n|^2 > |h_c|^2 + \tau\right). \end{aligned} \tag{6.37}$$

Similar to mode 1, the pairing condition correlates both $|h_c|^2$ and $|h_n|^2$. Therefore, conditioned on $|h_c|^2$, all eligible nCHs of mode 0 have the following CDF:

$$F_{|h_n|^2}^0(h_n) = \begin{cases} 0, & \text{for } h_n \leq h_c + \tau \\ \frac{e^{-h_c-\tau}-e^{-h_n}}{e^{-h_c-\tau}}, & \text{for } h_n > h_c + \tau. \end{cases} \tag{6.38}$$

Furthermore, the corresponding PDF for the eligible nCHs, denoted by $f_{|h_n|^2}^0(h_n)$, can be obtained as:

$$f_{|h_n|^2}^0(h_n) = \frac{e^{-h_n}}{e^{-h_c-\tau}}, \qquad \text{for } h_n > h_c + \tau. \tag{6.39}$$

Using mathematical manipulations, and considering the case of $\varepsilon > 1$, the outage probability in

(6.37) can be expressed as:

$$P_{o,P}^{nCH,0} = e^{-Q_0} - \frac{e^{\tau-\varepsilon_t}}{\varepsilon} e^{-\varepsilon Q_0}, \qquad \varepsilon > 1, \tag{6.40}$$

where $Q_0 = (0, q)^+$, and $q = (\tau - \varepsilon_t)/(\varepsilon - 1)$. Note that, for the cases of $\varepsilon \leq 1$, the outage probability either equal to zero for $\tau > \varepsilon_t$ or otherwise has very small value that can be neglected, specifically when τ is very close to ε_t. Using (6.36) and (6.40), the wasted resources for NOMA-mode 0 with active CH can be obtained as in (6.33).

In the same context, to complete the analysis while having active CH and nCH pairs, the CH outage with pair is calculated as follows:

$$\begin{aligned} P_{o,P}^{CH,0} =& 1 - \mathbb{P}\Big(\log(1 + P_t|h_c|^2) \geq R_{min}, \log(1 + \frac{P_t|h_n|^2}{P_t|h_c|^2 + 1}) \geq R_{min}, |h_n|^2 - |h_c|^2 \geq \tau\Big) \\ =& 1 - \mathbb{P}\Big(|h_c|^2 \geq \varepsilon_t, |h_n|^2 \geq \varepsilon(|h_c|^2 + \frac{1}{P_t}), |h_n|^2 \geq |h_c|^2 + \tau\Big). \end{aligned} \tag{6.41}$$

Using (6.39), the joint pdf $f^0_{|h_n|^2,|h_c|^2}$ can be used to obtain $P_{o,P}^{CH,0}$ as follows:

$$P_{o,P}^{CH,0} = \begin{cases} 1 - e^{-\varepsilon_t} + e^{-Q} - \frac{e^{\tau-\varepsilon_t-\varepsilon Q}}{\varepsilon}, & \varepsilon > 1, \\ 1 - e^{-Q} - \frac{e^{\tau-\varepsilon_t}}{\varepsilon}\left(e^{-\varepsilon\varepsilon_t} - e^{-\varepsilon Q}\right), & \varepsilon \leq 1, \end{cases} \tag{6.42}$$

where $Q = (\varepsilon_t, q)^+$.

Regarding the pairing success probability with active CH in mode 0, one needs to find $\mathbb{P}(N_{g_{c_i}} \neq 0)$. Given that $|h_1|^2 < |h_2|^2 < \cdots < |h_{N_{c_i}}|^2$, and using the pairing condition of mode 0 defined in (6.17), we have:

$$\mathbb{P}(N_{g_{c_i}} = n_g) = \mathbb{P}(|h_{N_{c_i}-n_g+1}|^2 \geq |h_c|^2 + \tau, |h_{N_{c_i}-n_g}|^2 < |h_c|^2 + \tau). \tag{6.43}$$

Given that $|h_{N_{c_i}}|^2$ has the pdf in (6.35), the probability of having at least one eligible nCH pair in mode 0 can be found as:

$$\begin{aligned} \mathbb{P}(N_{g_{c_i}} \neq 0) =& 1 - \mathbb{P}(|h_{N_{c_i}}|^2 < |h_c|^2 + \tau) \\ =& 1 - \frac{e^{\tau}}{N_{c_i} + 1}\left(1 - (1 - e^{-\tau})^{N_{c_i}+1}\right). \end{aligned} \tag{6.44}$$

Table 6.2: System model parameters

Parameter	Definition	Value
M	Number of uplink resources (RBs)	20
N	Number of MTDs	5000
N_a	Number of active MTDs	700:500:2700
T	Total number of cycles	10^4, 2×10^4
P_M, P_F, σ^2	Parameters of activity predictor	0.01, 0.01, 0.04
P_t	Transmission power	10:40 dBm
P_{tol}	Detection threshold for SIC	4 dBm
ρ	Weight for nCH reward division	0.2

6.4 Simulations

In the simulations, we consider $N = 5000$ MTDs randomly located at fixed points of a square area with side length 500 meters. Initially, the activation of the MTDs follows the Beta distribution with parameters ($\alpha = 3, \beta = 4$) within $I_A = 10$ time slots. Afterwards, we assume N_a randomly selected MTDs are active per cycle. Note that by considering cases of $N_a >> M$, the overload situation sustains.

Unless otherwise stated, the following parameters are used in the simulations. The parameters of the utility function (5.6) are set as $\delta_1 = 0.1, \delta_2 = 0.2, \delta_3 = 0.7$, where the delay is the major reward factor. Additionally, we set the parameters of the function in (5.7) as $a = 1, b = 8, c = 0.05$. To better illustrate the ability of MAB learning in achieving the essential delay requirement of MTC, we chose to set the first $N/2$ MTDs to be with strict maximum delay of $D_i \in [1, 20]$ slots, whereas the remaining MTDs are with relaxed maximum delay of $D_i \in [60, 100]$ slots. Running the simulations for a total of T cycles, the maximum delay of the MTDs changes with every activation. Furthermore, the predictor outputs a probability of being active P_a for each device that is selected randomly in the range $[0.8, 1]$. Table 6.2 shows the parameters used in the simulations.

6.4.1 Resource Wastage and Outage Results

Figure 6.2 plots the probability of selecting inactive MTDs p_i while increasing N_a where both simulations and analysis (6.5) are matching. As expected, p_i is inversely proportional with N_a and is not a function of the minimum rate R_{min} or transmission powers P_t. In addition, Figure 6.3 illustrates the average number of active MTDs per cluster in NOMA where both

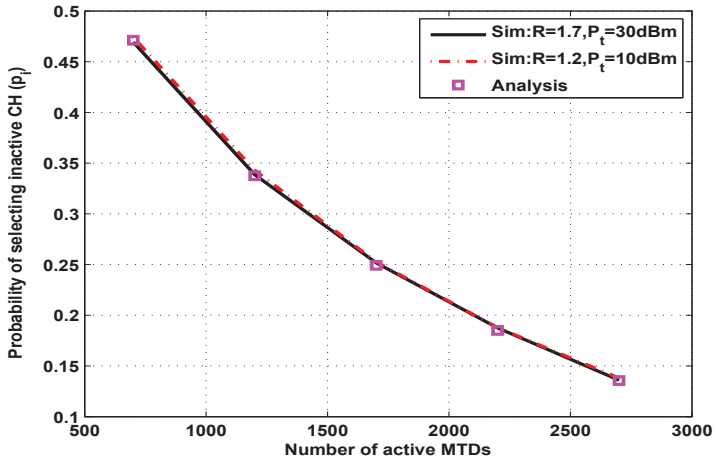

Figure 6.2: The probability of selecting inactive CH using fast uplink grant vs the number of active MTDs with $N = 5000$

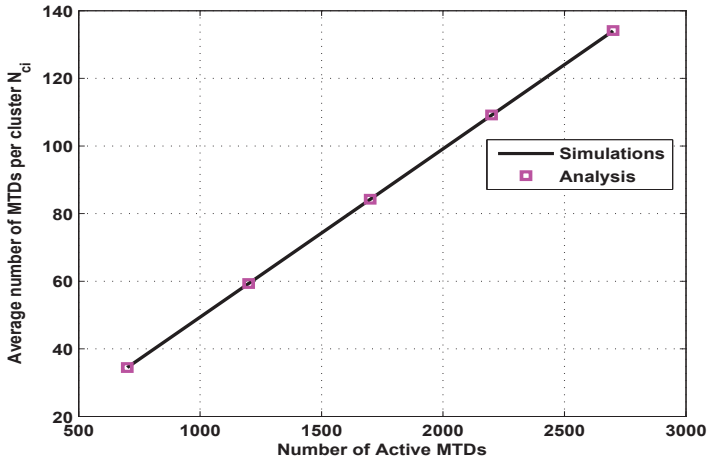

Figure 6.3: The average number of active MTDs per cluster N_{c_i} in NOMA vs the number of active MTDs with $N = 5000$

simulations and analysis are in perfect match. Using (6.12), the average $N_{c_i} = E\{N_{ar}\}/M$ for the Binomial distrubtion.

Considering $N_a = 1700$, the outages of OMA and NOMA scenarios while increasing P_t are depicted in the following figures. Starting by mode 1 in NOMA case, Figure 6.4-(a) shows that the simulation and analytical results of CH outage without a pair are identical. It also shows that increasing the transmission power reduces the outage probability as the CH transmits solely. It is worth pointing out that OMA MTDs' outages have similar curves to that shown in Figure 6.4-(a). On the other hand, Figure 6.4-(b) illustrates the CH outage with pair and verifies the accuracy of our analytical results under different values of R_{min}, P_{tol}. Unlike the no-pair

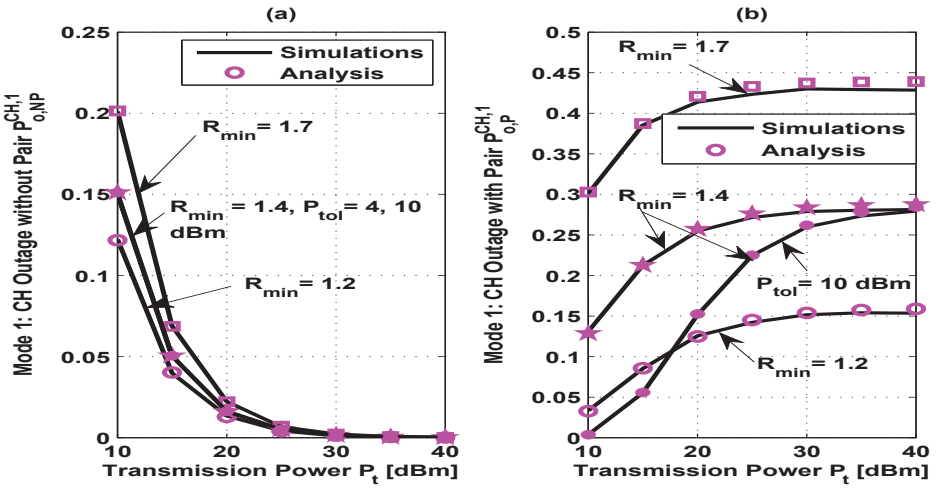

Figure 6.4: NOMA mode 1: (a) CH outage probability without a pair found $P_{o,NP}^{CH,1}$ (b) CH outage probability with pair found $P_{o,P}^{CH,1}$. The lines are the simulations and the markers are the analysis.

case, the results in Figure 6.4-(b) show that increasing P_t results in an increase in the outage probability. Since equal power allocation is assumed, increasing P_t is translated into tougher interference to the strong MTD pair (i.e. CH in mode 1) which increases the outage. This impact can be mitigated by elevating the threshold required for pairing by forcing higher P_{tol}. As the figure depicts for the same rate requirement of $R_{min} = 1.4$, using $P_{tol} = 10$ dBm enhances the outage performance, specifically at low-power range. Although increasing P_{tol} may be restrictive and limits the number of available pairs, the massive number of MTDs in IoT networks can compensate for this limitation such that eligible pairs could always exist.

Continuing in NOMA mode 1, the nCH outage with both inactive and active CH pair are illustrated in Figure 6.5-(a), (b), respectively. As seen, both simulations and analytical results are in very good agreement. In addition, the results clarify the enhancement achieved in the outage probability for the nCH (i.e. weak MTD) with the increase of the transmission power. However, in case of active CH pair, the outage probability of the nCH is larger than that of the corresponding CH pair due to SIC error propagation. It shall be emphasized that the outage results in Figure 6.5-(b) will have no contribution in the resource wastage probability.

Considering the above results of p_i and outages, the resources wastage for both OMA and NOMA systems are plotted in Figure 6.6. Except for the case of the restrictive rate requirement $R_{min} = 1.7$ where the MTDs outage is high, NOMA systems succeed to save the spectrum resources as compared to OMA. Even more, the proposed NOMA scheme is able to go beyond the minimum OMA wastage limit identified by p_i by allowing nCH to exploit the grants allo-

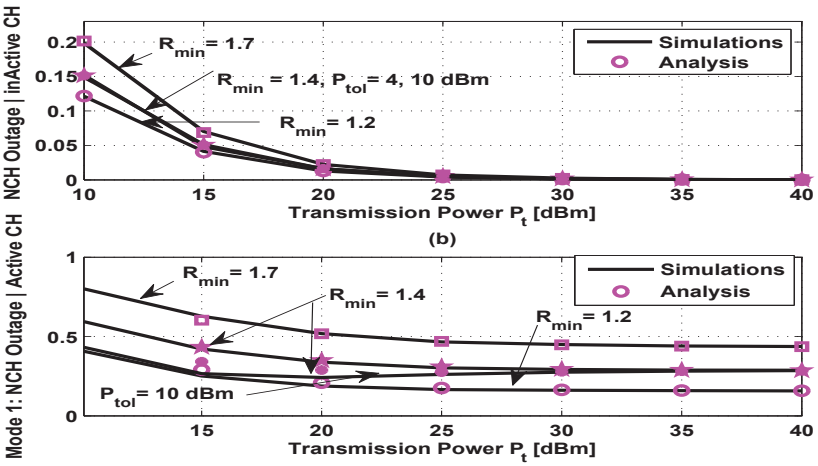

Figure 6.5: NOMA mode 1: (a) nCH outage probability given paired with inactive CH $P_o^{nCH}|$inactive CH (b) nCH outage probability given paired with active CH $P_{o,P}^{nCH,1}$. The lines are the simulations and the markers are the analysis.

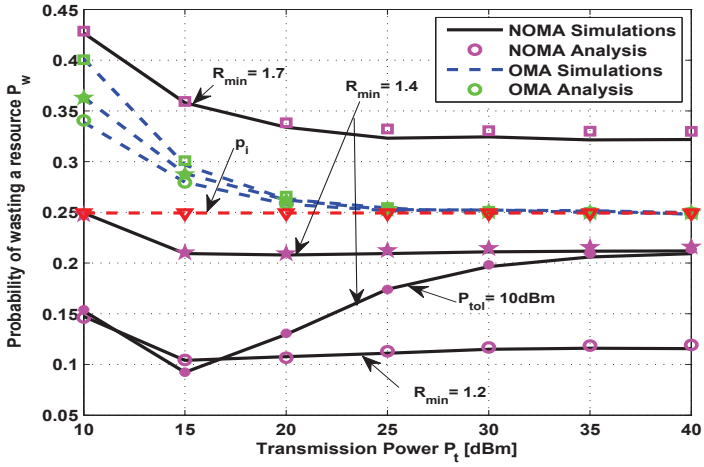

Figure 6.6: A comparison between the probability of resource wastage for both OMA and NOMA (mode 1) under different parameter configuration. For NOMA, lines are simulations and markers are analytical results. The red line is the probability of selecting inactive CH p_i.

cated to inactive MTDs. Although the outage probability may limit the NOMA performance, it is shown that by strengthening the pairing threshold (i.e. $P_{tol} = 10$ dBm) the outage can be reduced for the same rate requirement which further reduces the resources wastage.

To further verify the accuracy of our analytical results, Figure 6.7 shows the wastage probability of the resources in NOMA while increasing the number of active MTDs N_a. With both simulations and analysis match, the figure emphasizes that for low number of active MTDs, NOMA is able to boost the resource utilization while OMA suffers from high wastage. However, with increasing N_a, scheduling inactive MTDs is not the major factor of wastage. Instead,

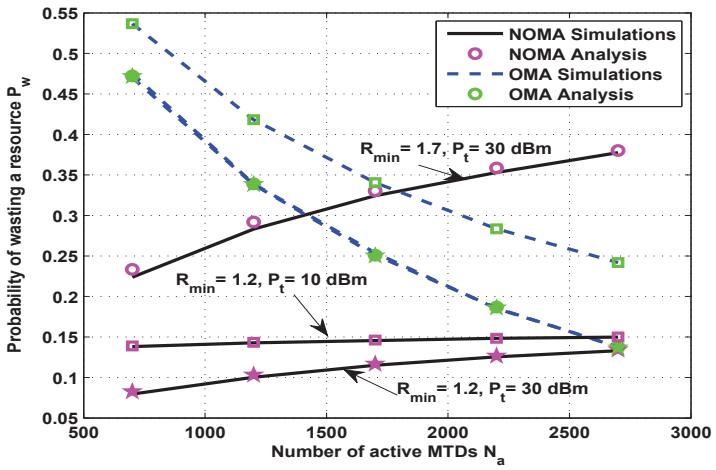

Figure 6.7: A comparison between the probability of resource wastage for both OMA and NOMA (mode 1) under different parameter configuration. Lines are simulations and markers are analytical results

outage probability dominates as the number of active MTDs increases. On the other hand, Figure 6.8 examines the effect of prediction error on both OMA and NOMA systems considering $N_a = 1700, R_{min} = 1.4, P_{tol} = 4$ dBm. Particularly, the figure shows that increasing the prediction error parameters definitely increases the probability of resource wastage for OMA. It can be seen that OMA wastage is lower-bounded by p_i. In contrast, the proposed NOMA systems attain significant enhancement in saving the resources via sharing them with other active MTDs (i.e. nCHs). Surprisingly, for the simulated system, increasing the error parameters results in a better utilization of the resources. The reason is that the simulated NOMA scenario is outage-limited and hence some resources is wasted due to SIC error propagation specifically at high power regime (due to equal power allocation). Consequently, the increase in the prediction error allows more nCHs to solely utilize more resources without sharing them. Definitely, utilizing the resources by suboptimal nCHs in the proposed NOMA is better than wasting the whole resource in case OMA is employed with high prediction error.

Considering NOMA mode 0, the outages of both paired CHs and nCHs are plotted in Figure 6.9-(a), (b), respectively. As mentioned in mode 1, Figure 6.9-(b) shows that the impact of interference is more noticeable on the outage of the strong MTDs (i.e. nCHs for mode 0) with the increase of transmission power and can be mitigated by increasing P_{tol}. In contrast, increasing P_t *mostly* enhances the performance of the weak MTD (i.e. CH in mode 0). However, CHs still suffer from the error propagation effect of SIC such that their outages are higher than their paired nCHs. It is worth mentioning that the outage results for the CHs with no pair are omitted as they approach zero for the simulated parameters. For the nCH with inactive pair,

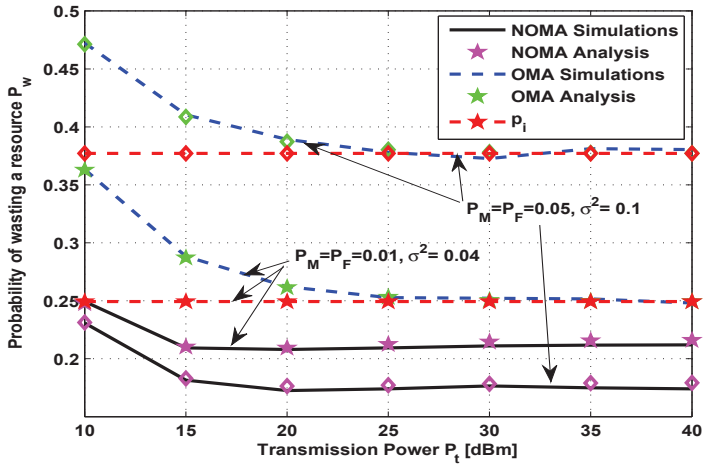

Figure 6.8: A comparison between the probability of resource wastage for both OMA and NOMA (mode 1) under different prediction error parameters. The red lines are the probability of selecting inactive CH p_i.

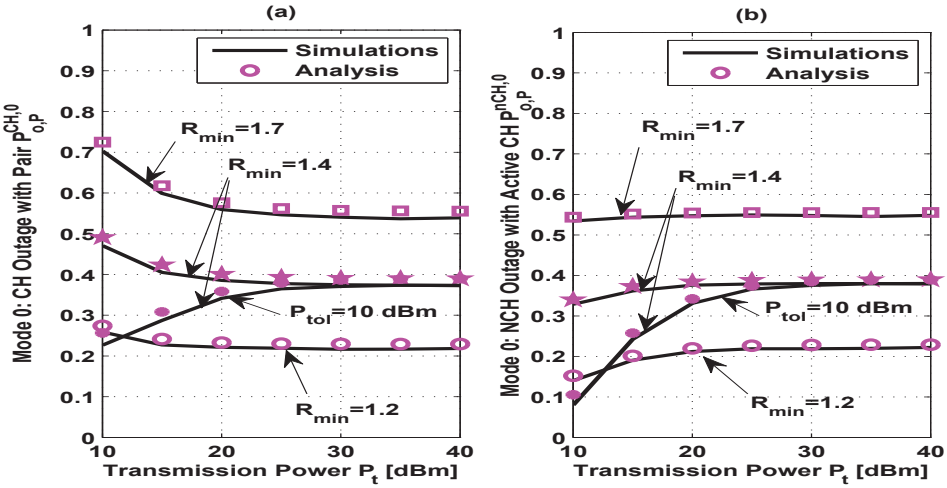

Figure 6.9: NOMA mode 0: (a) CH outage probability with pair $P_{o,P}^{CH,0}$ (b) nCH outage probability given paired with active CH $P_{o,P}^{nCH,0}$. The lines are the simulations and the markers are the analysis.

these results are mode independent and hence are the same as Figure 6.5-(a).

The resource wastage for mode 0 is demonstrated in Figure 6.10 where the matching between the developed analysis and simulations is clear. Additionally, the figure shows that for relatively low rate requirement (i.e. $R_{min} = 1.2$), NOMA noticeably reduces the resource wastage. However, as R_{min} and P_t simultaneously increase, the resource wastage of NOMA can slightly exceed OMA due to the increased outages of the MTDs. Strengthening the pairing threshold is shown to better tackle the outage problem and save more resources as in the case of $P_{tol} = 10$ dBm. It is worth mentioning that the assumption of equal power transmission can also be relaxed and power back-off can be used to better show the advantages of NOMA over

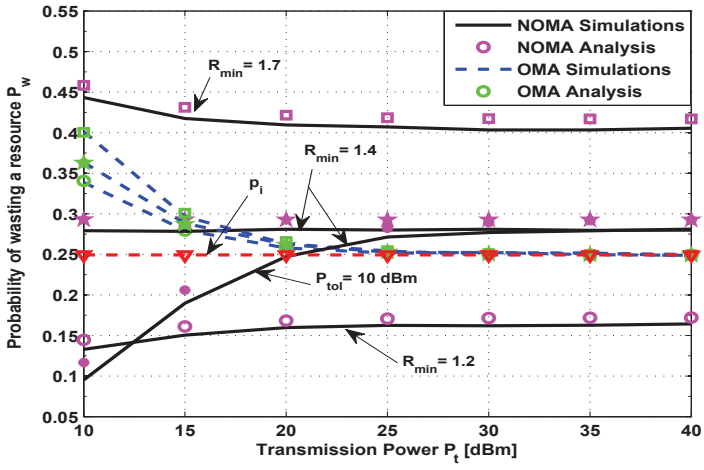

Figure 6.10: A comparison between the probability of resource wastage for both OMA and NOMA (mode 0) under different parameter configuration. For NOMA, lines are simulations and markers are analysis results. The red line is the probability of selecting inactive CH p_i.

OMA.

Power Back-off Allocation Strategy

The following figures examine the impact of applying a simple power back-off scheme on the outages and the resource wastage of NOMA mode 1. Particularly, we assume that in case a pair is admitted to the same RB with the active CH, the weak MTD will back-off its transmit power by a factor $\psi = 0.7$. Hence, in mode 1, we assume that the active CH transmit power is $P_s = P_t$, and the paired nCH transmit power is $P_w = \psi P_t$. Based on that the simulation results of the outages of both active CHs and nCHs while being paired with each other are demonstrated in Figure 6.11, and Figure 6.12, respectively. The figures show that the power back-off scheme is capable of boosting the outage performance of the MTDs and enhance the pairing experience. Particularly, all the MTDs achieve lower outage probability by adopting power back-off scheme as compared to the equal power allocation adopted in the previous results. In the same context, Figure 6.13 illustrates the simulation results of the resources wastage probability of both OMA and NOMA scenarios under the proposed power back-off allocation strategy as compared to the equal power allocation. The figure depicts the enhancement achieved in NOMA scenario by simple power back-off in saving the resources due to limiting the outages and mitigating the SIC error propagation. The same gains are also expected for NOMA mode 0. It is worth mentioning that these results are considered as initial results aiming to emphasize the potential enhancement that can by attained from the proposed NOMA-based fast grant scheme by further

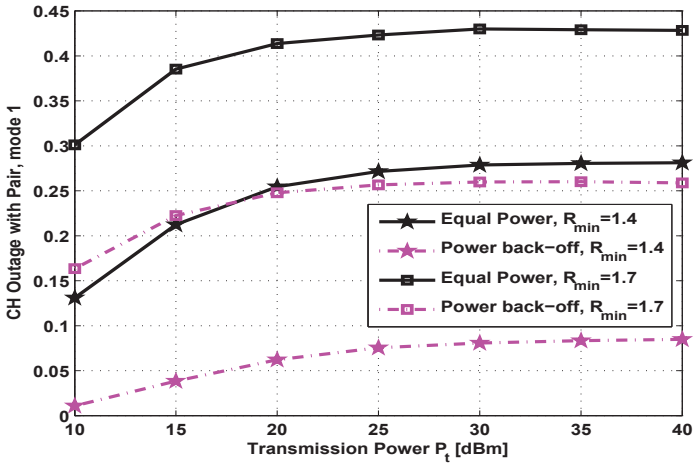

Figure 6.11: A comparison between the CH outage probability with pair in NOMA mode 1 assuming both equal power and power back-off allocations ($N_a = 1700$, $P_{tol} = 4$)

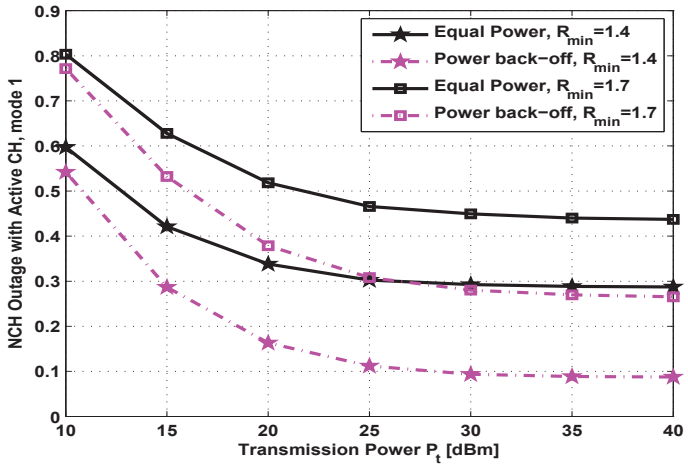

Figure 6.12: A comparison between the nCH outage probability with pair in NOMA mode 1 assuming both equal power and power back-off allocations ($N_a = 1700$, $P_{tol} = 4$)

enhancing the power allocation scheme.

6.4.2 Regret and Delay Results

In the following, P_{tol} is increased to 10 dBm to be able to pair more distinctive MTDs such that the interference impact on the strong MTD is mitigated. Also, we assume $N_a = 1700$ MTDs, $P_t = 20$ dBm for all MTDs and $R_{min} = 1.4$. Figure 6.14 plots the regrets of both OMA and NOMA scenarios where the reference reward is OMA-based attained by scheduling the set of M active MTDs with the highest rewards (i.e. best M MTDs). The figure shows the enhanced performance of NOMA where it achieves significantly lower regret as compared to its coun-

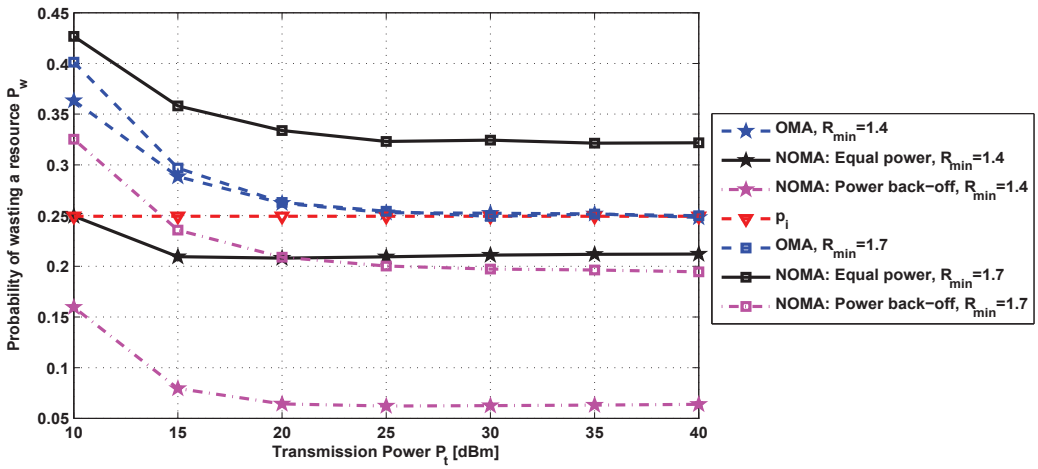

Figure 6.13: A comparison between the probability of resource wastage for both OMA and NOMA (mode 1) under different parameter configuration ($N_a = 1700, P_{tol} = 4$)

terpart OMA. Moreover, the figure shows that applying MAB learning for the CHs scheduling achieves extra enhancement for NOMA system as compared to random CHs scheduling. As seen, the gap between MAB and RND schemes is larger in NOMA as compared to the gap in OMA. This reflects that NOMA is able to better utilize the MAB learning benefits by serving more MTDs in less time which speeds the learning process. It shall be emphasized that in both NOMA-MAB and NOMA-RND the nCHs are randomly selected from an eligible set as described in Section 5.5.3.

These results are also shown in terms of the rewards in Figure 6.15-(a) where the reference reward is depicted in red. In addition, Figure 6.15-(b) shows the rewards of both OMA and NOMA systems while relaxing the target rate $R_{min} = 0.5$. The figure illustrates that the proposed NOMA scheme can attain or even exceeds the best OMA performance in terms of achieved rewards at an affordable complexity. For $R_{min} = 0.5$, the NOMA rewards exceed double the OMA reward not only by allowing more MTDs to share the resources but also by exploiting the resources wasted in OMA due to source traffic prediction errors. The rewards figures also reflect that OMA performance is limited mainly due to prediction errors as there is no big difference in the results while changing R_{min}. In contrast, NOMA is outage-limited, so that the relaxing of R_{min} results in significant improvement in the system. This can be interpreted that the proposed NOMA scheme is more robust against source traffic prediction errors and that the goal of decoupling the prediction and the scheduling performance is achieved. It shall be noted that NOMA mode 1 shows better performance than mode 0 for high rate requirements $R_{min} = 1.4$, and the reverse for the case of $R_{min} = 0.5$.

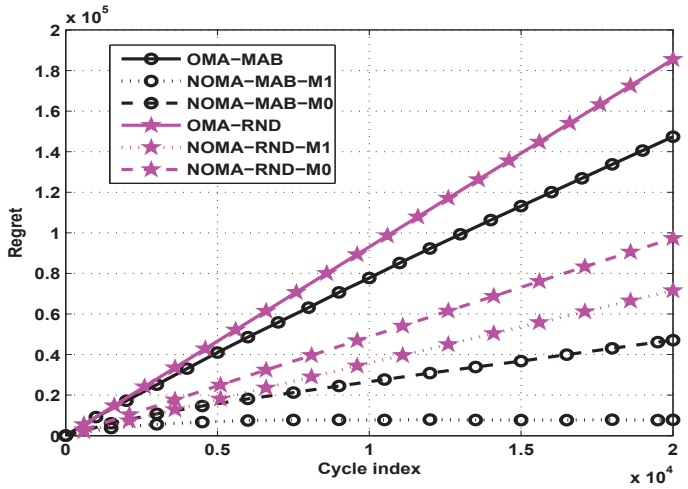

Figure 6.14: Regret of both OMA and NOMA scenarios where the reference reward is the best OMA MTDs assuming $R_{min} = 1.4$. For NOMA, M1, M0 refer to modes 1,0, respectively.

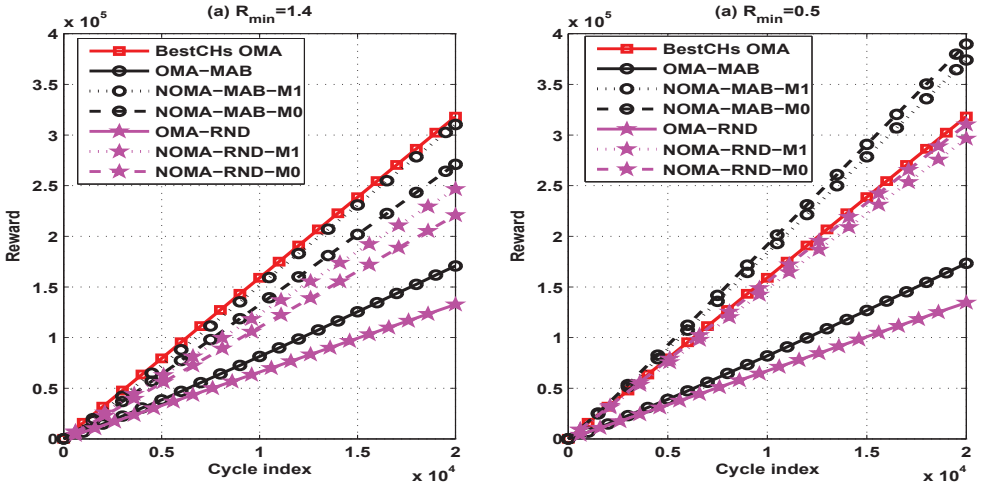

Figure 6.15: Rewards of both OMA and NOMA scenarios where the reference reward is the best OMA MTDs. For NOMA, M1, M0 refer to modes 1,0, respectively.

In the following, to better illustrate the enhancement achieved by MAB learning as compared to random CH scheduling, the scheduling histograms are given for mode 1 with $R_{min} = 1.4$. Figure 6.16-(a), (b) depict that MAB learning is able to select quality MTDs that have strict delay requirements (i.e the first $N/2$ MTDs) more frequently in both OMA and NOMA scenarios whereas NOMA scheduling frequency is higher (i.e. higher service rate). For NOMA, since MAB learning is only used for CHs scheduling the histogram in Figure 6.16-(c) shows that CHs are mostly MTDs with strict delay whereas Figure 6.16-(d) shows the uniform scheduling of the nCH MTDs regardless of their requirements (i.e. random nCHs selection). This approaches more the exploitation vs exploration balance known for learning algorithms by offering better

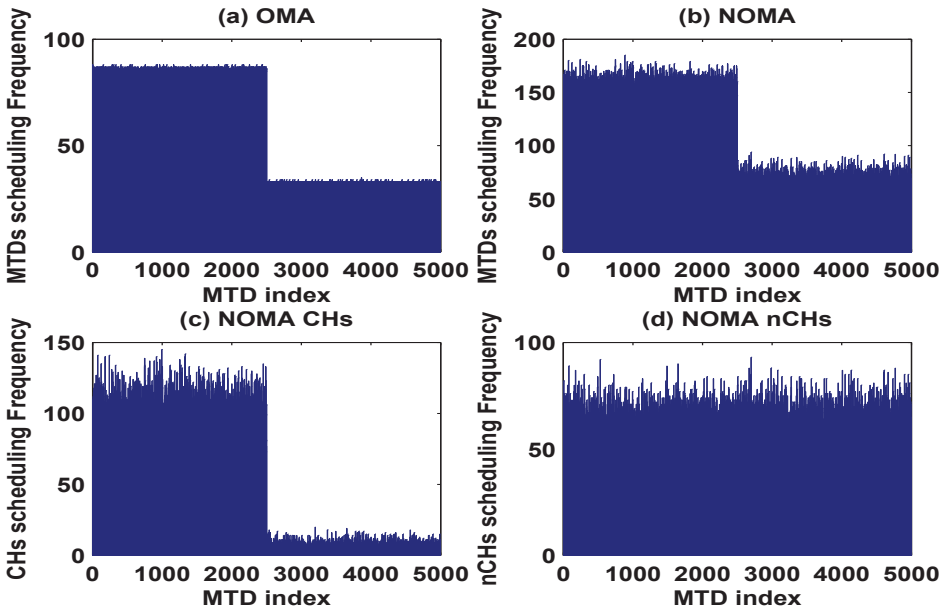

Figure 6.16: Histograms for the number of times the MTDs were scheduled of both OMA and NOMA scenarios with MAB learning for CHs

chance for MTDs with lower expected rewards to be selected among the nCHs. The corresponding histograms of totally random scheduling for both CHs and nCHs are shown in Figure 6.17.

Due to this difference in the selection distribution of the served MTDs, this simulated scenario shows that the proposed MAB-based NOMA can schedule MTDs achieving an average maximum tolerable delay around 31 cycles. Also, the CHs only achieve an average maximum tolerable delay around 14 cycles almost half of that achieved by OMA. In contrast, random NOMA for both CHs and nCHs achieves average maximum tolerable delay around 45 cycles. That is the proposed MAB-based NOMA allows faster access to more MTDs having more strict delay requirements. Table 6.3 summarizes the average maximum tolerable delay of the scheduled MTDs for different studied cases whereas the corresponding number of served MTDs, out of $MT = 4 \times 10^5$ available allocation opportunities, are depicted in the shaded entries of the table. It shall be noted that the little increase in the average maximum delay of MAB-based NOMA as compared to OMA implies that the number of served nCHs (i.e. randomly selected MTDs) is slightly higher than CHs. Consequently, with enhanced pairing, the aggregate average maximum tolerable delay can be noticeably improved. The same results are shown in Table 6.4 for the case of relaxed $R_{min} = 0.5$ where the number of served MTDs with NOMA is significantly higher than OMA, in addition to the enhanced delay performance.

It shall be emphasized that the proposed NOMA scheme provides flexibility to tailor the

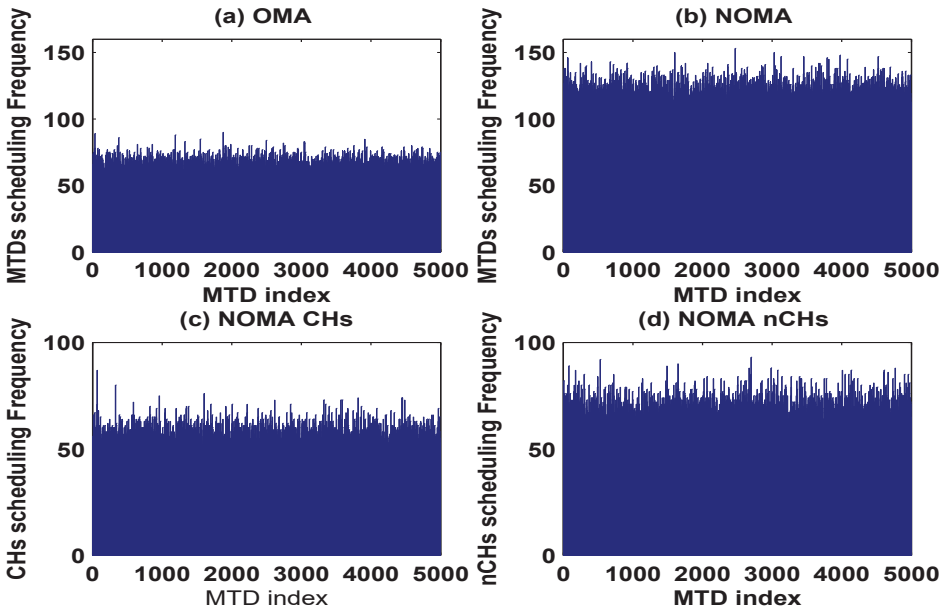

Figure 6.17: Histograms for the number of times the MTDs were scheduled of both OMA and NOMA scenarios with random selection of the CHs

Table 6.3: Average maximum delay of the selected MTDs with $R_{min} = 1.4$

		OMA	**NOMA**	
			Mode 1	**Mode 0**
MAB	*Aggregate*	29.6563	31.5522	33.2422
		29.4765×10^4	54.9767×10^4	49.2331×10^4
	CHs only		14.3966	14.3046
			24.9569×10^4	19.7583×10^4
Random	*Aggregate*	45.2258	45.2950	45.3222
		29.5328×10^4	54.9682×10^4	49.2595×10^4
	CHs only		45.3249	45.4482
			24.9629×10^4	19.6921×10^4

* Shaded entries are the corresponding number of served MTDs.

distinction among the served MTDs for given network performance parameters. This can be done via adapting the utility weights in (5.6) and latency mapping parameters in (5.7), in addition to the minimum level of distinction defined by P_{tol}. In the previous results, we showed the superiority of NOMA with MAB as compared to OMA as well as the random selection of the CHs. For the same given network variables of $N_a = 1700, P_t = 20$ dBm and the restriction of $R_{min} = 1.4$, we would have ended to less gains with the following selection of the proposed scheme parameters. Consider the case where the utility function (5.6) parameters are set as $\delta_1 = 0.2, \delta_2 = 0.3, \delta_3 = 0.5$. In addition, we set the parameters of the function in (5.7) as $a = 1, b = 8, c = 0.03$. In addition to $P_{tol} = 4$ dBm. In this case, the distinction among

Table 6.4: Average maximum delay of the selected MTDs with $R_{min} = 0.5$

		OMA	NOMA	
			Mode 1	Mode 0
MAB	*Aggregate*	29.5441	31.3097	31.8756
		29.8945×10^4	66.0010×10^4	69.3275×10^4
	CHs only		13.6343	13.3773
			29.8742×10^4	29.9404×10^4
Random	*Aggregate*	45.1437	45.2902	45.2760
		29.8999×10^4	66.0631×10^4	69.1945×10^4
	CHs only		45.2920	45.3524
			29.9752×10^4	29.8243×10^4

* Shaded entries are the corresponding number of served MTDs.

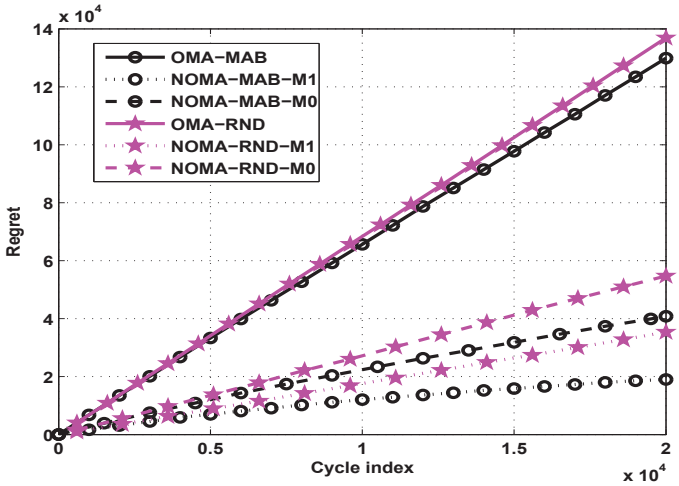

Figure 6.18: Regret of both OMA and NOMA scenarios where the reference reward is the best OMA MTDs. For NOMA, M1, M0 refer to modes 1,0, respectively.

the MTDs of strict latency and service priority as compared to the MTDs that able to tolerate higher delays is not very high. In other words, the rewards of scheduling an MTD either with strict delay or not are close in values. Hence, the benefits of MAB may not be noticed and it may seem similar to random scheduling. The results of this scenario are shown in Figure 6.18, and Figure 6.19. Although as before, the proposed NOMA MAB is better than others, the gap in the performance is reduced due to the lack of the distinction between the two groups of MTDs. Note also that this selection of the parameters affects the best system reward taken as a reference.

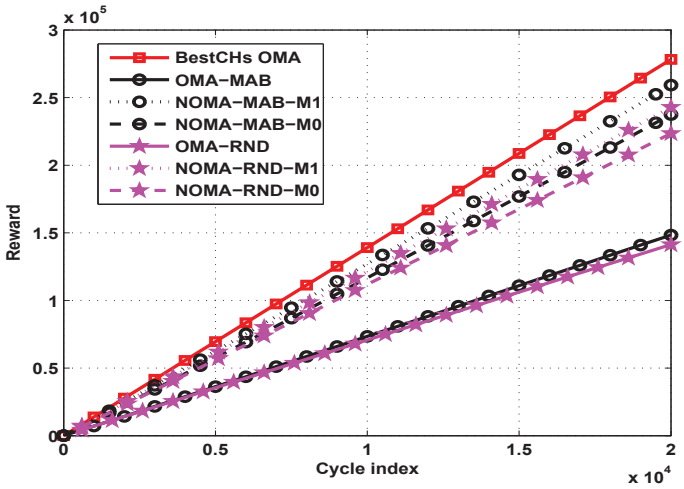

Figure 6.19: Rewards of both OMA and NOMA scenarios where the reference reward is the best OMA MTDs. For NOMA, M1, M0 refer to modes 1,0, respectively.

6.5 Summary

We have introduced a two-stage uplink access scheme for MTC that combines both fast uplink grant and NOMA. MAB learning was adopted for the BS to schedule the fast grant MTDs with no prior information about QoS or CSI. Then, NOMA was employed such that the selected MTDs can pick randomly selected pairs under certain constrains in a distributed manner with limited information exchange. In this chapter, we devised an abstraction model for the source traffic predictor needed for fast grant such that the prediction error can be evaluated. Consequently, the performance of the proposed scheme is analyzed in terms of average resource wastage, and outage probability. The simulation results show the effectiveness of the proposed method in saving the scarce resources while verifying the analysis accuracy. These results reflect the robustness of the proposed scheme against source traffic prediction errors. Moreover, the simulation results show the effectiveness of the proposed technique in providing access to more MTDs in a timely manner while accommodating their heterogeneous QoS requirements. However, it was shown that the proposed NOMA scheme is outage-limited. Therefore, the adaption of the network parameters such as the SIC distinction level can mitigate the interference impact and reduce the outage probabilities. Consequently, the performance can be enhanced. We believe that this work is a first step illustrating the potential gains of NOMA with fast uplink grant with more enhancements that could be achieved in future works with more emphasis on pairing and power allocation.

7. Conclusions and Future Work

7.1 Conclusions

The diverse requirements of the M2M applications have pushed different standardization bodies, such as the 3GPP, towards specifying the requirements and challenges of MTC. M2M communication brings billions of devices, usually with small amount of data transmission, that should be connected to support different applications. This massive access causes network congestion and overload problems if not properly managed. Although, cellular networks have been identified as promising radio access technology for MTC, the traditional random access protocols used in Long Term Evolution-Advanced (LTE-A) are inefficient for MTC. For instance, for an enormous number of devices attempting to access the network simultaneously, or even in a short period, the number of collisions could increase rapidly. These frequent collisions lead to energy and resource wastage, packet loss and severe access delay which are all undesired for energy-constrained MTC devices. Hence, efficient management of MTC massive access becomes a must. In this thesis, we tackled the problem of the efficient uplink access of MTC specifically with massive connectivity aiming to mitigate network congestion. Our research has led to significant contributions, that is recognizable from the record of publications resulted from this work. Next, we briefly summarize those accomplishments.

In Chapter 3, we addressed the overload control problem arises due to the concurrent transmission of enormous number of MTDs. In this regard, we adopted the Beta distribution to model the MTDs' arrival in congestion cases as suggested by the 3GPP. Then, we proposed a network overload algorithm, referred to M2M-OSA, that is capable of allocating the resources to massive number of MTDs in a distributed manner with reduced overhead. Being able to allocate the resources within relatively-small, bounded contention time, M2M-OSA can achieve full resource utilization and save the MTDs' energy. Further enhancements where attained by allowing fast data transmission right after gaining successful access which is more suitable for

the M2M applications with small-sized data packets. For proper operation of the proposed algorithm, we devised a method for the estimation of the number of active MTDs at each time slot, which is neither known by the devices nor the BS. Furthermore, we discussed a number of non-ideal situations that could be faced in real-life deployment of M2M-OSA including the effect of misestimation and feedback errors along with possible solutions.

In Chapter 4, we theoretically assessed the performance of the proposed M2M-OSA. With the help of the triangular approximation, the Beta-distributed traffic, known to be mathematically intractable, was analyzed. Particularly, closed-form expressions were provided for different evaluation metrics including average access delay, total service time, and energy consumption considering both lightly, and seriously congested cases. Moreover, the blocking probability was mathematically evaluated for the case of finite retransmission limit. The problem was then formulated as a binary integer programming problem and we showed that proposed algorithm approaches the optimal access delay at lower complexity. Finally, an updated version of M2M-OSA was provided to include a traffic regulation mechanism for further energy savings.

In Chapter 5, we studied the uplink access problem of the MTC considering both categories; mMTC and URLLC. Aiming to reduce latency as well as the signaling overhead, and to avoid the collisions of both coordinated and uncoordinated access, fast uplink grant concept was adopted. In fast uplink grant, the BS directly allocate the resources to the active MTDs without receiving a scheduling request which increases the risk of resources wastage in case an inactive MTD was selected. In our work, this risk was tackled by employing NOMA with fast uplink grant such that multiple MTDs can be allocated the same resource. Particularly, we proposed a two-stage uplink access technique that employs MAB learning to allocate the resources to the MTDs using fast uplink grant with no information about the MTDs' QoS or CSI at the BS. Afterwards, the scheduled MTDs seek pairing to other nearby active MTDs to share the uplink resources. To keep the advantage of limiting the information exchange, distributed pairing is proposed while maintaining minimum distinction among the MTDs for proper SIC operation. As a benchmark, a quasi-optimal NOMA scenario is formulated using BIP to assess the proposed scheme performance and illustrate the potential gains of NOMA with fast grant.

In Chapter 6, we provide an analytical framework for the probability of resources wastage in fast uplink grant considering both OMA and the proposed NOMA scheme. This was facilitated by devising an abstraction model for the prediction error of the traffic predictor that should be employed at the BS preceding the scheduling stage. Consequently, we provide mathematical

expressions for the average wastage of the resources as well as the outage probabilities while accounting for the MTDs' activity and SIC decoding order. The results emphasized the superiority of the proposed NOMA-based fast grant scheme in enhancing the resources utilization, and the MTDs' latency. Moreover, we shed light on the benefits of the proposed NOMA scheme in achieving better balance between exploration and exploitation trade-off while being more robust against the source traffic prediction errors. This trade-off was handled by allowing MAB to schedule the critical MTDs more frequently whereas other MTDs can be allowed to share the same resources in a NOMA fashion.

7.2 Future Work

In our work, we studied the uplink access of MTDs with a specific focus on congestion cases. For the proposed M2M-OSA, we assumed that all the devices have equivalent access opportunities. However, the inclusion of the different QoS requirements of the MTDs in the selection process is highly recommended as a future direction. This can be done by modifying the competing metric employed by the MTDs to reflect certain desired criterion that is application-dependent such as the latency or the throughput. Furthermore, different groups of MTDs with various priority levels could be assumed in the system which facilitates the inclusion of alarm traffic.

Although Beta distribution is a good representative for congested scenarios as specified by the 3GPP, the investigation of other types of traffics while employing the proposed M2M-OSA is another future direction. This definitely affects the analysis procedure related to both the average access delay and the total service time that depends on the triangle approximation of the Beta distribution. In contrast, we believe that this modification should has no effect on the analyses of the energy consumption and the blocking probability that are dependent on the M2M-OSA procedure itself. In addition, a formal study for the non-ideal situations discussed earlier related to the proposed M2M-OSA is recommended as future work along with the relaxation of the ideal channel condition assumption.

For the NOMA-based fast grant work, we employed equal power allocation and random selection of the nCH pair. Since our work of NOMA with fast grant is among the very first trials of combining both schemes, we focused on showing the potential gains of this combination while employing simplified pairing and power allocation techniques. The BIP formulation showed

the expected enhancements that could be further attained by enhancing the pairing technique. For instance, instead of random pairing, we may allow the CH to be paired to the eligible nCH having the poorest or the best channel conditions in mode 1 and mode 0, respectively. In such case, the level of distinction between the MTDs is maximized which reduces the probability of outage. On the other hand, more sophisticated power allocation techniques such as enhanced power back-off with optimized back-off factor can also be utilized to mitigate the interference suffered by the strong MTD and reduce the outage probability.

Another promising future direction is to allow the selection of the nCH MTD pair, not only based on the channel conditions, but formalizing a utility that combines different criteria depending on the application. For instance, we may allow the CH to seek pairing to nCHs that has maximum utility combines both its channel condition and the delay tolerance. The utility should also be tailored to comply with the pairing mode either best or poorest channel conditions. To keep the pairing distributed, OSA can also be utilized among other distributed schemes such as the timer-based contention control mechanism proposed in [136]. Additionally, extending the learning-based scheduling employed on our work to select all the scheduled MTDs while adopting NOMA is encouraged as a future work. Finally, increasing the NOMA cluster size (i.e more than two devices) and applying other decoding techniques would be also an excellent future direction.

Appendix A:
List of Publications

Journal Articles

[J4] **M. E. Tanab** and W. Hamouda, "Efficient Resource Allocation in Fast Uplink Grant for Machine-Type Communications With NOMA," *IEEE Internet of Things Journal,* June 2021, Submitted.

[J3] **M. El-Tanab** and W. Hamouda, "An Overview of Uplink Access Techniques in Machine-Type Communications," *IEEE Network,* Nov. 2020.

[J2] **M. E. Tanab** and W. Hamouda, "Machine-to-Machine Communications With Massive Access: Congestion Control," *IEEE Internet of Things Journal,* vol. 6, no. 2, pp. 3545-3557, April 2019.

[J1] **M. E. Tanab** and W. Hamouda, "Resource Allocation for Underlay Cognitive Radio Networks: A survey," *IEEE Communications Surveys & Tutorials,* vol. 19, no. 2, pp. 1249-1276, Secondquarter 2017.

Conference Proceedings

[C3] **M. E. Tanab** and W. Hamouda, "Fast Grant Learning-Based Approach for Machine Type Communications with NOMA," *IEEE International Conference on Communications (ICC'21), Montreal, Canada,* 2021, Accepted. **(Best Paper Award)**

[C2] **M. E. Tanab** and W. Hamouda, "Performance of Overload Control in Machine- to-Machine Wireless Networks," *IEEE Global Communications Conference (GLOBECOM), Abu Dhabi, United Arab Emirates,* 2018, pp. 1-7.

[C1] **M. E. Tanab,** and W. Hamouda, "A Scalable Overload Control Algorithm for Massive Access in Machine-to-Machine Networks," *IEEE International Conference on Communications (ICC'17), Paris, France,* May 2017, pp. 1-6.

Appendix B:
Outage Proofs of The Proposed NOMA-Based Fast Grant Scheme: MODE 1

CH Outage Without a Pair

$$\begin{aligned} P_{o,NP}^{CH,1} =& \mathbb{P}\left(\log(1+P_t|h_c|^2) < R_{min}, |h_1|^2 > |h_c|^2 - \tau\right) \\ =& \mathbb{P}\left(|h_c|^2 < \varepsilon_t, |h_c|^2 < |h_1|^2 + \tau\right), \end{aligned} \tag{1}$$

For the case of $\tau \geq \varepsilon_t$

$$\begin{aligned} P_{o,NP}^{CH,1} =& \mathbb{P}\left(|h_c|^2 < \varepsilon_t\right) \\ =& \int_0^{\infty}\int_0^{\varepsilon_t} N_{c_i} e^{-N_{c_i}h_1} e^{-h_c} \mathrm{d}h_c \mathrm{d}h_1 = \int_0^{\infty} N_{c_i} e^{-N_{c_i}h_1}\left(1-e^{-\varepsilon_t}\right)\mathrm{d}h_1 \\ =& 1 - e^{-\varepsilon_t} \qquad\qquad \text{for } \tau \geq \varepsilon_t. \end{aligned} \tag{2}$$

For the case of $\tau < \varepsilon_t$

$$\begin{aligned} P_{o,NP}^{CH,1} =& \int_0^{\varepsilon_t-\tau}\int_0^{h_1+\tau} N_{c_i} e^{-N_{c_i}h_1} e^{-h_c} \mathrm{d}h_c \mathrm{d}h_1 + \int_{\varepsilon_t-\tau}^{\infty}\int_0^{\varepsilon_t} N_{c_i} e^{-N_{c_i}h_1} e^{-h_c} \mathrm{d}h_c \mathrm{d}h_1 \\ =& I_1 + I_2, \end{aligned} \tag{3}$$

The integration I_1 is evaluated as follows:

$$\begin{aligned} I_1 =& \int_0^{\varepsilon_t-\tau} N_{c_i} e^{-N_{c_i}h_1}(1-e^{-h_1-\tau})\mathrm{d}h_1 \\ =& 1 - e^{-N_{c_i}(\varepsilon_t-\tau)} - \frac{N_{c_i}e^{-\tau}}{1+N_{c_i}}(1-e^{-(1+N_{c_i})(\varepsilon_t-\tau)}) \end{aligned}$$

The integration I_2 is evaluated as follows:

$$\begin{aligned} I_2 =& \int_{\varepsilon_t-\tau}^{\infty} N_{c_i} e^{-N_{c_i} h_1} (1-e^{-\varepsilon_t}) \mathrm{d}h_1 \\ =& (1-e^{-\varepsilon_t}) e^{-N_{c_i}(\varepsilon_t-\tau)}. \end{aligned}$$

Substituting I_1 and I_2 in (3), then:

$$P_{o,NP}^{CH,1} = 1 - \frac{N_{c_i} e^{-\tau}}{1+N_{c_i}} \left(1 - e^{-(1+N_{c_i})(\varepsilon_t-\tau)}\right) - e^{-N_{c_i}(\varepsilon_t-\tau)-\varepsilon_t}, \qquad \text{for } \tau < \varepsilon_t. \tag{4}$$

CH Outage With Pair

$$\begin{aligned} P_{o,P}^{CH,1} =& \mathbb{P}\left(\log(1+\frac{P_t|h_c|^2}{P_t|h_n|^2+1}) < R_{min}, |h_c|^2 - |h_n|^2 > \tau\right) \\ =& \mathbb{P}\left(|h_c|^2 < \varepsilon(|h_n|^2 + \frac{1}{P_t}), |h_c|^2 > |h_n|^2 + \tau\right). \end{aligned} \tag{5}$$

Considering the case of $\varepsilon > 1$, and assuming $\tau > \varepsilon_t$:

$$P_{o,P}^{CH,1} = \int_q^{\infty} \int_{h_n+\tau}^{\varepsilon h_n+\varepsilon_t} \frac{e^{-h_n} e^{-h_c}}{1-e^{\tau-h_c}} \, \mathrm{d}h_c \, \mathrm{d}h_n$$

where $\varrho = \tau - \varepsilon_t$, and $q = \varrho/(\varepsilon - 1)$. Consider the following substitution:

$$\text{let } u = 1 - e^{\tau-h_c}, \mathrm{d}u = e^{\tau} e^{-h_c}$$

Then, $P_{o,P}^{CH,1}$ is expressed as follows:

$$\begin{aligned} P_{o,P}^{CH,1} =& \int_q^{\infty} e^{-h_n} \int \frac{e^{-\tau}}{u} \, \mathrm{d}u \, \mathrm{d}h_n \\ =& \int_q^{\infty} e^{-\tau} e^{-h_n} \left[\ln|1 - e^{\tau} e^{-\varepsilon h_n - \varepsilon_t}| - \ln|1 - e^{-h_n}|\right] \\ =& e^{-\tau} \left[I_3 - I_4\right] \end{aligned} \tag{6}$$

The integration I_4 is evaluated as follows:

$$I_4 = \int_q^{\infty} e^{-h_n} \ln|1 - e^{-h_n}| \, \mathrm{d}h_n$$

Using integration by parts, let

$$\begin{aligned} u =& \ln|1-e^{-h_n}| & \mathrm{d}v =& e^{-h_n} \\ \mathrm{d}u =& \frac{e^{-h_n}}{1-e^{-h_n}}\,\mathrm{d}h_n & v =& -e^{-h_n} \end{aligned}$$

Based on that I_4 is equivalent to:

$$\begin{aligned} I_4 =& e^{-h_n}\ln|1-e^{-h_n}|\Big|_{\infty}^{q} + \int_q^{\infty}\frac{e^{-2h_n}}{1-e^{-h_n}}\,\mathrm{d}h_n \\ =& e^{-q}\ln|1-e^{-q}| + I_{41} \end{aligned} \tag{7}$$

Consider the substitution of $k = 1-e^{-h_n}, \mathrm{d}k = e^{-h_n}\,\mathrm{d}h_n$, then I_{41} is evaluated as:

$$\begin{aligned} I_{41} =& \int\frac{1-k}{k}\,\mathrm{d}k = \ln|k| - k \\ =& \left[\ln|1-e^{-h_n}| - (1-e^{-h_n})\right]_q^{\infty} \\ =& -\left(e^{-q} + \ln|1-e^{-q}|\right) \end{aligned} \tag{8}$$

Substituting (8) into (7), we get:

$$I_4 = -\left[(1-e^{-q})\ln|1-e^{-q}| + e^{-q}\right]. \tag{9}$$

Back to (6), the integration I_3 is evaluated considering $\varrho = \tau - \varepsilon_t$ as follows:

$$I_3 = \int_q^{\infty} e^{-h_n}\ln|1-e^{\varrho}e^{-\varepsilon h_n}|\,\mathrm{d}h_n$$

Using integration by parts, let

$$\begin{aligned} u =& \ln|1-e^{\varrho}e^{-\varepsilon h_n}| & \mathrm{d}v =& e^{-h_n} \\ \mathrm{d}u =& \frac{\varepsilon e^{\varrho}e^{-\varepsilon h_n}}{1-e^{\varrho}e^{-\varepsilon h_n}}\,\mathrm{d}h_n & v =& -e^{-h_n} \end{aligned}$$

Based on that I_3 is equivalent to:

$$\begin{aligned} I_3 =& \left[e^{-h_n} \ln \left| 1 - e^{\varrho} e^{-\varepsilon h_n} \right| \right]_{\infty}^{q} + \int_{q}^{\infty} \frac{\varepsilon e^{\varrho} e^{-h_n(\varepsilon+1)}}{1 - e^{\varrho} e^{-\varepsilon h_n}} \, \mathrm{d}h_n \\ =& e^{-q} \ln \left| 1 - e^{\varrho} e^{-\varepsilon q} \right| + I_{31} \end{aligned} \tag{10}$$

Consider the substitution of $k = 1 - e^{\varrho} e^{-\varepsilon h_n}, \mathrm{d}k = \varepsilon e^{\varrho} e^{-\varepsilon h_n} \, \mathrm{d}h_n$, then I_{31} is evaluated as:

$$I_{31} = \int_{k=1-e^{\varrho}e^{-\varepsilon q}}^{1} \frac{e^{-h_n}}{k} \, \mathrm{d}k = \int_{k=1-e^{\varrho}e^{-\varepsilon q}}^{1} e^{-\frac{\varrho}{\varepsilon}} \frac{(1-k)^{\frac{1}{\varepsilon}}}{k} \, \mathrm{d}k$$

Consider also the substitution of $t = (1-k)^{\frac{1}{\varepsilon}}$, then $k = 1 - t^{\varepsilon}, \mathrm{d}k = -\varepsilon t^{\varepsilon-1} \mathrm{d}t$. Based on that, I_{31} is evaluated as:

$$I_{31} = \int e^{-\frac{\varrho}{\varepsilon}} \frac{-t}{1-t^{\varepsilon}} \, \varepsilon t^{\varepsilon-1} \, \mathrm{d}t = \int_{t=0}^{e^{\frac{\varrho}{\varepsilon}} e^{-q}} e^{-\frac{\varrho}{\varepsilon}} \frac{\varepsilon t^{\varepsilon}}{1-t^{\varepsilon}} \mathrm{d}t$$

Using integration by parts, let

$$\begin{aligned} u =& t & \mathrm{d}v =& \frac{\varepsilon t^{\varepsilon-1}}{1-t^{\varepsilon}} \\ \mathrm{d}u =& \mathrm{d}t & v =& -\ln \left| 1 - t^{\varepsilon} \right| \end{aligned}$$

Based on that I_{31} is evaluated as:

$$I_{31} = e^{-\frac{\varrho}{\varepsilon}} \left(-e^{\frac{\varrho}{\varepsilon}} e^{-q} \ln \left| 1 - e^{\varrho} e^{-q\varepsilon} \right| + \int_{0}^{e^{\frac{\varrho}{\varepsilon}} e^{-q}} \ln \left| 1 - t^{\varepsilon} \right| \, \mathrm{d}t \right) \tag{11}$$

Substituting (11) into (10), we get I_3 as follows:

$$I_3 = e^{-\frac{\varrho}{\varepsilon}} \int_{0}^{e^{\frac{\varrho}{\varepsilon}} e^{-q}} \ln \left| 1 - t^{\varepsilon} \right| \, \mathrm{d}t \tag{12}$$

Substituting I_3 and I_4 in (6), then $P_{o,P}^{CH,1}$ for $\varepsilon > 1, \tau > \varepsilon_t$ is expressed as:

$$P_{o,P}^{CH,1} = e^{-\tau} \left(e^{-\frac{\varrho}{\varepsilon}} \int_{0}^{e^{\frac{\varrho}{\varepsilon}} e^{-q}} \ln \left| 1 - t^{\varepsilon} \right| \, \mathrm{d}t + (1 - e^{-q}) \ln \left| 1 - e^{-q} \right| + e^{-q} \right) \tag{13}$$

For the case of $\tau \leq \varepsilon_t$, set $q = 0$ as follows:

$$L_1 = \lim_{q \to 0} (1 - e^{-q}) \ln |1 - e^{-q}| \tag{14}$$

Since L_1 is undefined, let $x = 1 - e^{-q})$. Then, we get:

$$L_1 = \lim_{x \to 0} x \ln |x| = \lim_{x \to 0} \frac{\ln |x|}{\frac{1}{x}}$$

Using L'Hôpital's rule:

$$L_1 = \lim_{x \to 0} \frac{\frac{1}{x}}{-\frac{1}{x^2}} = \lim_{x \to 0} -x = 0 \tag{15}$$

Hence, $P_{o,P}^{CH,1}$ for $\varepsilon > 1, \tau \leq \varepsilon_t$ is expressed as:

$$P_{o,P}^{CH,1} = e^{-\tau} \left(e^{-\frac{\varrho}{\varepsilon}} \int_0^{e^{\frac{\varrho}{\varepsilon}}} \ln |1 - t^{\varepsilon}| \, \mathrm{d}t + 1 \right) \qquad , \varepsilon > 1, \tau \leq \varepsilon_t. \tag{16}$$

NCH Outage With Pair

$$\begin{aligned} P_{o,P}^{nCH,1} =& 1 - \mathbb{P}\Big(\log(1 + P_t |h_n|^2) \geq R_{min}, \log(1 + \frac{P_t |h_c|^2}{P_t |h_n|^2 + 1}) \geq R_{min}, |h_c|^2 - |h_n|^2 > \tau \Big) \\ =& 1 - \mathbb{P}\Big(|h_n|^2 \geq \varepsilon_t, |h_c|^2 \geq \varepsilon(|h_n|^2 + \frac{1}{P_t}), |h_c|^2 > |h_n|^2 + \tau \Big). \end{aligned} \tag{17}$$

Define $Q = (q, \varepsilon_t)^+$ such that $q = (\tau - \varepsilon_t)/(\varepsilon - 1)$ as before. Based on that, $P_{o,P}^{nCH,1}$ is obtained as:

$$\begin{aligned} P_{o,P}^{nCH,1} =& 1 - \int_Q^{\infty} \int_{\varepsilon h_n + \varepsilon_t}^{\infty} \frac{e^{-h_n} e^{-h_c}}{1 - e^{\tau - h_c}} \, \mathrm{d}h_c \, \mathrm{d}h_n - \int_{\varepsilon_t}^{Q} \int_{h_n + \tau}^{\infty} \frac{e^{-h_n} e^{-h_c}}{1 - e^{\tau - h_c}} \, \mathrm{d}h_c \, \mathrm{d}h_n \\ =& 1 - I_5 - I_6 \end{aligned} \tag{18}$$

To evaluate I_5, let $k = 1 - e^{\tau - h_c}, \mathrm{d}k = e^{\tau} e^{-h_c}\, \mathrm{d}h_c$. Hence, we get:

$$\begin{aligned} I_5 &= \int_Q^{\infty} \int e^{-h_n} \frac{e^{-\tau}}{k}\, \mathrm{d}k = -\int_Q^{\infty} e^{-\tau} e^{-h_n} \ln|1 - e^{-\varepsilon h_n} e^{\tau - \varepsilon_t}|\, \mathrm{d}h_n \\ &= -e^{-\tau} I_3 = -e^{-\tau} e^{-\frac{\varrho}{\varepsilon}} \int_0^{e^{\frac{\varrho}{\varepsilon}} e^{-Q}} \ln|1 - t^{\varepsilon}|\, \mathrm{d}t \end{aligned} \tag{19}$$

Following the same steps, I_6 is evaluated as:

$$I_6 = -\int_{\varepsilon_t}^{Q} e^{-\tau} e^{-h_n} \ln|1 - e^{-h_n}|\, \mathrm{d}h_n$$

Using integration by parts, let

$$\begin{aligned} u &= \ln|1 - e^{-h_n}| & \mathrm{d}v &= -e^{-h_n} \\ \mathrm{d}u &= \frac{e^{-h_n}}{1 - e^{-h_n}}\, \mathrm{d}h_n & v &= e^{-h_n} \end{aligned}$$

Based on that I_6 is expressed as:

$$\begin{aligned} I_6 &= e^{-\tau} \left(e^{-Q} \ln|1 - e^{-Q}| - e^{-\varepsilon_t} \ln|1 - e^{-\varepsilon_t}| - \int_{\varepsilon_t}^{Q} \frac{e^{-2h_n}}{1 - e^{-h_n}}\, \mathrm{d}h_n \right) \\ &= e^{-\tau} \left(e^{-Q} \ln|1 - e^{-Q}| - e^{-\varepsilon_t} \ln|1 - e^{-\varepsilon_t}| - I_{61} \right) \end{aligned} \tag{20}$$

Using the substitution of $k = 1 - e^{-h_n}, \mathrm{d}k = e^{-h_n}\, \mathrm{d}h_n$, I_{61} is as follows:

$$\begin{aligned} I_{61} &= \int_{1 - e^{-\varepsilon_t}}^{1 - e^{-Q}} \frac{(1 - k)}{k} \mathrm{d}k = \Big[\ln|k| - k \Big]_{1 - e^{-\varepsilon_t}}^{1 - e^{-Q}} \\ &= \ln|1 - e^{-Q}| + e^{-Q} - \ln|1 - e^{-\varepsilon_t}| - e^{-\varepsilon_t} \end{aligned} \tag{21}$$

Substituting (21) in (20), we have I_6 as follows:

$$\begin{aligned} I_6 &= e^{-\tau} \Big((e^{-Q} - 1) \ln|1 - e^{-Q}| + (1 - e^{-\varepsilon_t}) \ln|1 - e^{-\varepsilon_t}| + e^{-\varepsilon_t} - e^{-Q} \Big) \\ &= c_1 \ln|c_1| - c_2 \ln|c_2| - c_1 + c_2 \end{aligned} \tag{22}$$

where $c_1 = 1 - e^{-\varepsilon_t}$, and $c_2 = 1 - e^{-Q}$. Substituting (19), and (22) into (18), $P_{o,P}^{nCH,1}$ can be easily obtained.

Pairing Probability

Given that the CH is active, we answer the question of " what is the probability of not finding a pair?".

We consider that at each cluster C_i there is N_{c_i} members where $N_{g_{c_i}}$ among them are eligible for pairing. Particularly, a nCH n is eligible to pair to CH c in mode 1 iff $|h_n|^2 \leq |h_c|^2 - \tau$. Considering that the channel gains of the cluster members are ordered as $|h_1|^2 < |h_2|^2 < \cdots < |h_{N_{c_i}}|^2$, then we have:

$$\mathbb{P}(N_{g_{c_i}} = n_g) = \mathbb{P}\Big(|h_{n_g}|^2 \leq |h_c|^2 - \tau, |h_{n_g+1}|^2 > |h_c|^2 - \tau\Big). \tag{23}$$

Based on that the probability of not finding a pair is equivalent to:

$$\begin{aligned}
\mathbb{P}(N_{g_{c_i}} = 0) =& \mathbb{P}\left(|h_1|^2 > |h_c|^2 - \tau\right) \\
=& \int_0^{\infty} \int_0^{h_1+\tau} N_{c_i} e^{-N_{c_i} h_1} e^{-h_c} \, \mathrm{d}h_c \mathrm{d}h_1 = \int_0^{\infty} N_{c_i} e^{-N_{c_i} h_1} (1 - e^{-h_1-\tau}) \, \mathrm{d}h_1 \\
=& N_{c_i} \left[\frac{e^{-N_{c_i} h_1}}{N_{c_i}} - e - \tau \frac{e^{-h_1(N_{c_i}+1)}}{N_{c_i}+1} \right]_{\infty}^{0} \\
=& 1 - \frac{N_{c_i} e^{-\tau}}{N_{c_i}+1}.
\end{aligned} \tag{24}$$

where $|h_1|^2 \sim \exp(N_{c_i})$.

Appendix C:
Outage Proofs of The Proposed NOMA-Based Fast Grant Scheme: MODE 0

CH Outage Without a Pair

$$\begin{aligned}
P_{o,NP}^{CH,0} =& \mathbb{P}\Big(\log(1+P_t|h_c|^2) < R_{min}, |h_{N_{c_i}}|^2 < |h_c|^2+\tau\Big) \\
=& \mathbb{P}\Big(|h_c|^2 < \varepsilon_t, |h_c|^2 > |h_{N_{c_i}}|^2-\tau\Big), \\
=& \int_0^{\varepsilon_t}\int_0^{h_c+\tau} N_{c_i}\left(1-e^{-h}\right)^{N_{c_i}-1} e^{-h}e^{-h_c}\,\mathrm{d}h\,\mathrm{d}h_c \\
=& \int_0^{\varepsilon_t} e^{-h_c}(1-e^{-h})^{N_{c_i}}\Big|_0^{h_c+\tau}\mathrm{d}h_c = \int_0^{\varepsilon_t} e^{-h_c}(1-e^{-h_c}e^{-\tau})^{N_{c_i}}\,\mathrm{d}h_c
\end{aligned}$$

Using the substitution $k = 1-e^{-h_c}e^{-\tau}, \mathrm{d}k = e^{-h_c}e^{-\tau}$, we get:

$$\begin{aligned}
P_{o,NP}^{CH,0} =& \int_0^{\varepsilon_t} e^{\tau}k^{N_{c_i}}\,\mathrm{d}k = \frac{e^{\tau}}{N_{c_i}+1}\left(1-e^{-h_c}e^{-\tau}\right)^{N_{c_i}+1}\Big|_0^{\varepsilon_t} \\
=& \frac{e^{\tau}}{N_{c_i}+1}\Big[(1-e^{-\tau-\varepsilon_t})^{N_{c_i}+1} - (1-e^{-\tau})^{N_{c_i}+1}\Big]. \qquad (25)
\end{aligned}$$

NCH Outage With Pair

$$\begin{aligned}
P_{o,P}^{nCH,0} =& \mathbb{P}\left(\log(1+\frac{P_t|h_n|^2}{P_t|h_c|^2+1}) < R_{min}, |h_n|^2-|h_c|^2 > \tau\right) \\
=& \mathbb{P}\left(|h_n|^2 < \varepsilon(|h_c|^2+\frac{1}{P_t}), |h_n|^2 > |h_c|^2+\tau\right). \qquad (26)
\end{aligned}$$

Considering the case of $\varepsilon > 1$, and assuming $\tau > \varepsilon_t$:

$$\begin{aligned} P_{o,P}^{nCH,0} =& \int_{q}^{\infty} \int_{h_c+\tau}^{\varepsilon h_c+\varepsilon_t} e^{\tau} e^{-h_n} \, \mathrm{d}h_n \, \mathrm{d}h_c \\ =& \int_{q}^{\infty} e^{\tau} \left(e^{-h_c-\tau} - e^{-\varepsilon h_c-\varepsilon_t} \right) \mathrm{d}h_c \\ =& e^{-q} - \frac{e^{\tau-\varepsilon_t}}{\varepsilon} e^{-\varepsilon q}, \qquad\qquad \varepsilon > 1, \tau > \varepsilon_t \end{aligned} \tag{27}$$

where again $\varrho = \tau - \varepsilon_t$, and $q = \varrho/(\varepsilon - 1)$. For the case of $\tau \leq \varepsilon_t$, we set $q = 0$ as follows:

$$P_{o,P}^{nCH,0} = 1 - \frac{e^{\tau-\varepsilon_t}}{\varepsilon}, \qquad\qquad \varepsilon > 1, \tau \leq \varepsilon_t \tag{28}$$

CH Outage With Pair

$$\begin{aligned} P_{o,P}^{CH,0} =& 1 - \mathbb{P}\left(\log(1 + P_t |h_c|^2) \geq R_{min}, \log(1 + \frac{P_t |h_n|^2}{P_t |h_c|^2 + 1}) \geq R_{min}, |h_n|^2 - |h_c|^2 \geq \tau \right) \\ =& 1 - \mathbb{P}\left(|h_c|^2 \geq \varepsilon_t, |h_n|^2 \geq \varepsilon(|h_c|^2 + \frac{1}{P_t}), |h_n|^2 \geq |h_c|^2 + \tau \right). \end{aligned} \tag{29}$$

Define $Q = (q, \varepsilon_t)^+$ such that $q = (\tau - \varepsilon_t)/(\varepsilon - 1)$ as before. Considering $\varepsilon > 1$, $P_{o,P}^{CH,0}$ is obtained as:

$$\begin{aligned} P_{o,P}^{CH,0} =& 1 - \int_{Q}^{\infty} \int_{\varepsilon h_c+\varepsilon_t}^{\infty} e^{\tau} e^{-h_n} \, \mathrm{d}h_n \, \mathrm{d}h_c - \int_{\varepsilon_t}^{Q} \int_{h_c+\tau}^{\infty} e^{\tau} e^{-h_n} \, \mathrm{d}h_n \, \mathrm{d}h_c \\ =& 1 - I_7 - I_8. \end{aligned} \tag{30}$$

First, we evaluate I_7 as:

$$I_7 = \int_{Q}^{\infty} e^{\tau} e^{-\varepsilon h_c-\varepsilon_t} \, \mathrm{d}h_c = \frac{e^{\tau-\varepsilon_t-\varepsilon Q}}{\varepsilon}. \tag{31}$$

Then, I_8 is evaluated as:

$$I_8 = \int_{\varepsilon_t}^{Q} e^{\tau} e^{-h_c-\tau} \, \mathrm{d}h_c = e^{-\varepsilon_t} - e^{-Q}. \tag{32}$$

Hence, $P_{o,P}^{CH,0}$ is expressed as follows:

$$P_{o,P}^{CH,0} = 1 - e^{-\varepsilon_t} + e^{-Q} - \frac{e^{\tau-\varepsilon_t-\varepsilon Q}}{\varepsilon}, \qquad \varepsilon > 1. \tag{33}$$

Next, we study the case of $\varepsilon \leq 1$. Starting by the case of $\tau > \varepsilon_t$, we get $P_{o,P}^{CH,0}$ as follows:

$$\begin{aligned} P_{o,P}^{CH,0} =& 1 - \int_{\varepsilon_t}^{\infty} \int_{h_c+\tau}^{\infty} e^{\tau} e^{-h_n} \, \mathrm{d}h_n \, \mathrm{d}h_c \\ =& 1 - \int_{\varepsilon_t}^{\infty} e^{\tau} e^{-h_c-\tau} \, \mathrm{d}h_c \\ =& 1 - e^{-\varepsilon_t}, \qquad \tau > \varepsilon_t, \varepsilon \leq 1. \end{aligned} \tag{34}$$

For the case of $\tau \leq \varepsilon_t$, $P_{o,P}^{CH,0}$ is evaluated as follows:

$$\begin{aligned} P_{o,P}^{CH,0} =& 1 - \int_{\varepsilon_t}^{Q} \int_{\varepsilon h_c+\varepsilon_t}^{\infty} e^{\tau} e^{-h_n} \, \mathrm{d}h_n \, \mathrm{d}h_c - \int_{Q}^{\infty} \int_{h_c+\tau}^{\infty} e^{\tau} e^{-h_n} \, \mathrm{d}h_n \, \mathrm{d}h_c \\ =& 1 - I_9 - I_{10}. \end{aligned} \tag{35}$$

I_9 is obtained as:

$$I_9 = \int_{\varepsilon_t}^{Q} e^{\tau} e^{-\varepsilon h_c-\varepsilon_t} \, \mathrm{d}h_c = \frac{e^{\tau-\varepsilon_t}}{\varepsilon} \Big(e^{-\varepsilon\varepsilon_t} - e^{-\varepsilon Q} \Big) \tag{36}$$

Also, I_{10} is obtained as:

$$I_{10} = \int_{Q}^{\infty} e^{\tau} e^{-h_c-\tau} \, \mathrm{d}h_c = e^{-Q}. \tag{37}$$

Finally, $P_{o,P}^{CH,0}$ is expressed as:

$$P_{o,P}^{CH,0} = 1 - e^{-Q} - \frac{e^{\tau-\varepsilon_t}}{\varepsilon} \Big(e^{-\varepsilon\varepsilon_t} - e^{-\varepsilon Q} \Big), \qquad \tau \leq \varepsilon_t, \varepsilon \leq 1. \tag{38}$$

Pairing Probability

In mode 0, a nCH n is eligible to pair to CH c in mode 1 iff $|h_n|^2 \geq |h_c|^2 + \tau$. Considering that the channel gains of the cluster members are ordered as $|h_1|^2 < |h_2|^2 < \cdots < |h_{N_{c_i}}|^2$, then we

have:

$$\mathbb{P}(N_{g_{c_i}} = n_g) = \mathbb{P}(|h_{N_{c_i}-n_g+1}|^2 \geq |h_c|^2 + \tau, |h_{N_{c_i}-n_g}|^2 < |h_c|^2 + \tau). \tag{39}$$

Based on that the probability of not finding a pair is equivalent to:

$$\begin{aligned}
\mathbb{P}(N_{g_{c_i}} = 0) =& \mathbb{P}(|h_{N_{c_i}}|^2 < |h_c|^2 + \tau) \\
=& \int_0^\infty \int_0^{h_c+\tau} N_{c_i} \left(1 - e^{-h}\right)^{N_{c_i}-1} e^{-h} e^{-h_c} \, \mathrm{d}h \, \mathrm{d}h_c \\
=& \int_0^\infty e^{-h_c} (1 - e^{-h})^{N_{c_i}} \Big|_0^{h_c+\tau} \mathrm{d}h_c \\
=& \int_0^\infty e^{-h_c} (1 - e^{-h_c} e^{-\tau})^{N_{c_i}} \mathrm{d}h_c
\end{aligned}$$

Using the substitution $k = 1 - e^{-h_c} e^{-\tau}, \mathrm{d}k = e^{-h_c} e^{-\tau}$, we get:

$$\begin{aligned}
\mathbb{P}(N_{g_{c_i}} = 0) =& \int_0^\infty e^{\tau} k^{N_{c_i}} \, \mathrm{d}k = \frac{e^\tau}{N_{c_i}+1} \left(1 - e^{-h_c} e^{-\tau}\right)^{N_{c_i}+1} \Big|_0^\infty \\
=& \frac{e^\tau}{N_{c_i}+1} \left(1 - (1 - e^{-\tau})^{N_{c_i}+1}\right).
\end{aligned} \tag{40}$$

References

[1] S. K. Sharma and X. Wang, "Toward Massive Machine Type Communications in Ultra-Dense Cellular IoT Networks: Current Issues and Machine Learning-Assisted Solutions," *IEEE Communications Surveys & Tutorials,* vol. 22, no. 1, pp. 426-471, Firstquarter 2020.

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

[3] 3GPP TS 22.368 V10.1.0, "Service requirements for machine-type communications," July 2010.

[4] M. Wang *et al.*, "Cellular machine-type communications: Physical challenges and solutions," *IEEE Wireless Communications,* vol. 23, no. 2, pp. 126-135, April 2016.

[5] I. Selinis, K. Katsaros, M. Allayioti, S. Vahid and R. Tafazolli, "The Race to 5G Era; LTE and Wi-Fi," *IEEE Access,* vol. 6, pp. 56598-56636, 2018.

[6] 3GPP TR 37.868 V11.0.0, "Study on RAN improvements for machine-type communications," Sep. 2011.

[7] R. Abbas, M. Shirvanimoghaddam, Y. Li and B. Vucetic,"A Novel Analytical Framework for Massive Grant-Free NOMA," *IEEE Transactions on Communications,* vol. 67, no. 3, pp. 2436-2449, March 2019.

[8] S. Duan, V. Shah-Mansouri, and V. W.S. Wong, "Dynamic access class barring for M2M communications in LTE networks," *In IEEE GLOBECOM Workshops (GC Wkshps),* Dec. 2013, pp. 4747-4752.

[9] F. M. Awuor, and C. Y. Wang, "Massive machine type communication in cellular system: A distributed queue approach," *IEEE International Conference on Communications (ICC), Kuala Lumpur,* May 2016, pp. 1-7.

[10] M. T. Islam, A. e. M. Taha, and S. Akl, "A survey of access management techniques in machine type communications," *IEEE Communications Magazine,* vol. 52, no. 4, pp. 74-81, April 2014.

[11] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan and V. K. Bhargava, "A Survey on Non-Orthogonal Multiple Access for 5G Networks: Research Challenges and Future Trends," *IEEE Journal on Selected Areas in Communications,* vol. 35, no. 10, pp. 2181-2195, Oct. 2017.

[12] B. Makki, K. Chitti, A. Behravan and M. -S. Alouini, "A Survey of NOMA: Current Status and Open Research Challenges," *IEEE Open Journal of the Communications Society,* vol. 1, pp. 179-189, 2020.

[13] W. Shin, M. Vaezi, B. Lee, D. J. Love, J. Lee and H. V. Poor, "Non-Orthogonal Multiple Access in Multi-Cell Networks: Theory, Performance, and Practical Challenges," *IEEE Communications Magazine,* vol. 55, no. 10, pp. 176-183, Oct. 2017.

[14] M. B. Shahab, R. Abbas, M. Shirvanimoghaddam and S. J. Johnson, "Grant-free Non-orthogonal Multiple Access for IoT: A Survey," *IEEE Communications Surveys & Tutorials,* doi: 10.1109/COMST.2020.2996032.

[15] S. Ali, N. Rajatheva and W. Saad, "Fast Uplink Grant for Machine Type Communications: Challenges and Opportunities," *IEEE Communications Magazine,* vol. 57, no. 3, pp. 97-103, March 2019.

[16] L. Dai, B. Wang, Z. Ding, Z. Wang, S. Chen and L. Hanzo, "A Survey of Non-Orthogonal Multiple Access for 5G," *IEEE Communications Surveys & Tutorials,* vol. 20, no. 3, pp. 2294-2323, thirdquarter 2018.

[17] M. E. Tanab, and W. Hamouda, "A scalable overload control algorithm for massive access in machine-to-machine networks," *IEEE International Conference on Communications (ICC'17), Paris, France,* May 2017, pp. 1-6.

[18] M. E. Tanab and W. Hamouda, "Performance of Overload Control in Machine- to-Machine Wireless Networks," *IEEE Global Communications Conference (GLOBECOM), Abu Dhabi, United Arab Emirates,* 2018, pp. 1-7.

[19] M. E. Tanab and W. Hamouda, "Machine-to-Machine Communications With Massive Access: Congestion Control," *IEEE Internet of Things Journal,* vol. 6, no. 2, pp. 3545-3557, April 2019.

[20] M. El-Tanab and W. Hamouda, "An Overview of Uplink Access Techniques in Machine-Type Communications," *IEEE Network,* doi: 10.1109/MNET.011.2000513.

[21] M. E. Tanab and W. Hamouda, "Fast Grant Learning-Based Approach for Machine Type Communications with NOMA," *IEEE International Conference on Communications (ICC'21), Montreal, Canada,* 2021, Accepted.

[22] M. E. Tanab and W. Hamouda, "Efficient Resource Allocation in Fast Uplink Grant for Machine-Type Communications With NOMA," *IEEE Internet of Things Journal,* June 2021, Submitted.

[23] G. A. Akpakwu, B. J. Silva, G. P. Hancke and A. M. Abu-Mahfouz, "A Survey on 5G Networks for the Internet of Things: Communication Technologies and Challenges," *IEEE Access,* vol. 6, pp. 3619-3647, 2018.

[24] M. I. Hossain, A. Laya, F. Militano, S. Iraji, and J. Markendahl, "Reducing signaling overload: Flexible capillary admission control for dense MTC over LTE networks," *IEEE 26th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), Hong Kong,* 2015, pp. 1305-1310.

[25] 3GPP TR 36.888 V12.0.0, "Study on Provision of Low-Cost Machine-Type Communications (MTC) User Equipments (UEs) Based on LTE," Sophia Antipolis, France, 2013. [Online]. Available: http://www.3gpp.org/ftp/

[26] M. Elsaadany, A. Ali and W. Hamouda, "Cellular LTE-A Technologies for the Future Internet-of-Things: Physical Layer Features and Challenges," *IEEE Communications Surveys & Tutorials,* vol. 19, no. 4, pp. 2544-2572, Fourthquarter 2017.

[27] H. Shariatmadari *et al.*, "Machine-type communications: current status and future perspectives toward 5G systems," *IEEE Communications Magazine,* vol. 53, no. 9, pp. 10-17, September 2015.

[28] M. Elbayoumi, M. Kamel, W. Hamouda and A. Youssef, "NOMA-Assisted Machine-Type Communications in UDN: State-of-the-Art and Challenges," *IEEE Communications Surveys & Tutorials,* vol. 22, no. 2, pp. 1276-1304, Secondquarter 2020.

[29] F. Ghavimi and H. H. Chen, "M2M communications in 3GPP LTE/LTE-A networks: Architectures, service requirements, challenges, and applications," *IEEE Communications Surveys & Tutorials,* vol. 17, no. 2, pp. 525-549, 2nd Quart., 2015.

[30] A. Rajandekar, and B. Sikdar, "A survey of MAC layer issues and protocols for machine-to-machine communications," *IEEE Internet of Things Journal,* vol. 2, no. 2, pp. 175-186, April 2015.

[31] A. Rico-Alvarino *et al.*, "An overview of 3GPP enhancements on machine to machine communications,"*IEEE Communications Magazine,* vol. 54, no. 6, pp. 14-21, June 2016.

[32] M. Hasan, E. Hossain and D. Niyato, "Random access for machine-to-machine communication in LTE-advanced networks: issues and approaches," *IEEE Communications Magazine,* vol. 51, no. 6, pp. 86-93, June 2013.

[33] M. Shirvanimoghaddam, Y. Li, M. Dohler, B. Vucetic, and S. Feng, "Probabilistic Rateless Multiple Access for Machine-to-Machine Communication," *IEEE Transactions on Wireless Communications,* vol. 14, no. 12, pp. 6815-6826, Dec. 2015.

[34] K. Zheng *et al.*, "Challenges of massive access in highly dense LTE-advanced networks with machine-to-machine communications," *IEEE Wireless Communications,* vol. 21, no. 3, pp. 12-18, June 2014.

[35] R. Ratasuk, A. Prasad, Z. Li, A. Ghosh and M. A. Uusitalo, "Recent advancements in M2M communications in 4G networks and evolution towards 5G," 18^{th} *International Conference on Intelligence in Next Generation Networks,* Paris, France, 2015, pp. 52-57.

[36] T. P. C. de Andrade, C. A. Astudillo and N. L. S. da Fonseca, "Allocation of Control Resources for Machine-to-Machine and Human-to-Human Communications Over LTE/LTE-A Networks," *IEEE Internet of Things Journal,* vol. 3, no. 3, pp. 366-377, June 2016.

[37] M. Wang *et al.*, "Cellular machine-type communications: physical challenges and solutions," *IEEE Wireless Communications,* vol. 23, no. 2, pp. 126-135, April 2016.

[38] A. Ali, W. Hamouda and M. Uysal, "Next generation M2M cellular networks: challenges and practical considerations," *IEEE Communications Magazine,* vol. 53, no. 9, pp. 18-24, September 2015.

[39] 3GPP TS 36.306, "E-UTRA, UE Radio Access Capabilities (Release 12, v.12.7.0)," 2015.

[40] D. Flore. (2015, Feb) Evolution of LTE in Release 13. [Online]. Available: http://www.3gpp.org/news-events/3gpp-news/1628-rel13

[41] M. Elsaadany and W. Hamouda, "The new enhancements in LTE-A Rel-13 for reliable machine type communications," *IEEE* 28^{th} *Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC),* Montreal, QC, Canada, 2017, pp. 1-5.

[42] A. Hoglund *et al.*, "Overview of 3GPP Release 14 Further Enhanced MTC," *IEEE Communications Standards Magazine,* vol. 2, no. 2, pp. 84-89, June 2018.

[43] 3GPP TR 36.881 V14.0.0, "Study on Latency Reduction Techniques for LTE," June 2016.

[44] C. Hoymann *et al.*, "LTE release 14 outlook," *IEEE Communications Magazine,* vol. 54, no. 6, pp. 44-49, June 2016.

[45] 3GPP R2-1709457, "Early Data Transmission Failure Handling in NB-IoT," 2017.

[46] "The 5G Evolution: 3GPP Releases 16-17," *5G Americas,* 2020.

[47] 3GPP RP-193264 "Rel-17 enhancements for NB-IoT and LTE-MTC," Dec. 2019.

[48] F. Montori, L. Bedogni, M. D. Felice, and L. Bononi, "Machine-to-machine wireless communication technologies for the Internet of Things: Taxonomy, comparison and open issues," *Pervasive Mobile Computing,* vol. 50, pp. 56–81, Oct. 2018.

[49] A. Čolaković and M. Hadžialić, "Internet of Things (IoT): A review of enabling technologies, challenges, and open research issues," *Computer Networks,* vol. 144, pp. 17–39, Oct. 2018.

[50] 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 Communications Surveys & Tutorials,* vol. 16, no. 1, pp. 4-16, First Quarter 2014.

[51] E. Soltanmohammadi, K. Ghavami and M. Naraghi-Pour, "A Survey of Traffic Issues in Machine-to-Machine Communications Over LTE," *IEEE Internet of Things Journal,* vol. 3, no. 6, pp. 865-884, Dec. 2016.

[52] L. Liu *et al.*, "Sparse Signal Processing for Grant-Free Massive Connectivity: A Future Paradigm for Random Access Protocols in the Internet of Things," *IEEE Signal Processing Magazine,* vol. 35, no. 5, pp. 88-99, Sept. 2018.

[53] M. Koseoglu, "Lower bounds on the LTE-A average random access delay under massive M2M arrivals," *IEEE Transactions on Communications,* vol. 64, no. 5, pp. 2104-2115, May 2016.

[54] M. Gerasimenko *et al.*, "Impact of machine-type communications on energy and delay performance of random access channel in LTE-advanced," *Trans. on Emerging Telecommun. Technol.,* vol. 24, no. 4, pp. 366-377, June 2013.

[55] A. Laya, L. Alonso, and J. Alonso-Zarate, "Efficient contention resolution in highly dense LTE networks for machine type communications," *IEEE Global Communications Conference (GLOBECOM), San Diego, CA,* Dec. 2015, pp. 1-7.

[56] D. T. Wiriaatmadja, and K. W. Choi, "Hybrid random access and data transmission protocol for machine-to-machine communications in cellular networks," *IEEE Transactions on Wireless Communications,* vol. 14, no. 1, pp. 33-46, Jan. 2015.

[57] D. J. Dechene and A. Shami, "Energy-Aware Resource Allocation Strategies for LTE Uplink with Synchronous HARQ Constraints," *IEEE Transactions on Mobile Computing,* vol. 13, no. 2, pp. 422-433, Feb. 2014.

[58] T. Kim, H. S. Jang and D. K. Sung, "An Enhanced Random Access Scheme With Spatial Group Based Reusable Preamble Allocation in Cellular M2M Networks," *IEEE Communications Letters,* vol. 19, no. 10, pp. 1714-1717, Oct. 2015.

[59] J. S. Kim, S. Lee and M. Y. Chung, "Efficient Random-Access Scheme for Massive Connectivity in 3GPP Low-Cost Machine-Type Communications," *IEEE Transactions on Vehicular Technology,* vol. 66, no. 7, pp. 6280-6290, July 2017.

[60] NTT DOCOMO, "Uplink multiple access schemes for NR," 3GPP TSG-RAN WG1 Meeting #85, Nanjing, China, document R1-165174, May 2016.

[61] H. Jiang *et al.*, "Distributed Layered Grant-Free Non-Orthogonal Multiple Access for Massive MTC," *IEEE* 29^{th} *Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC),* Bologna, 2018, pp. 1-7.

[62] S. Ali, A. Ferdowsi, W. Saad, N. Rajatheva and J. Haapola, "Sleeping Multi-Armed Bandit Learning for Fast Uplink Grant Allocation in Machine Type Communications," *IEEE Transactions on Communications,* vol. 68, no. 8, pp. 5072-5086, Aug. 2020.

[63] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication.* Cambridge, U.K.: Cambridge Univ. Press, 2005.

[64] Y. Zhang, K. Peng, Z. Chen and J. Song, "SIC vs. JD: Uplink NOMA techniques for M2M random access," *IEEE International Conference on Communications (ICC),* Paris, 2017, pp. 1-6.

[65] K. Higuchi and A. Benjebbour, "Non-orthogonal multiple access (NOMA) with successive interference cancellation for future radio access," *IEICE Trans. Commun.,* vol. E98-B, no. 3, pp. 403-414, Mar. 2015.

[66] Z. Yang, Z. Ding, P. Fan and N. Al-Dhahir, "A General Power Allocation Scheme to Guarantee Quality of Service in Downlink and Uplink NOMA Systems," *IEEE Transactions on Wireless Communications,* vol. 15, no. 11, pp. 7244-7257, Nov. 2016.

[67] O. Maraqa, A. S. Rajasekaran, S. Al-Ahmadi, H. Yanikomeroglu and S. M. Sait, "A Survey of Rate-Optimal Power Domain NOMA With Enabling Technologies of Future Wireless Networks," *IEEE Communications Surveys & Tutorials,* vol. 22, no. 4, pp. 2192-2235, Fourthquarter 2020.

[68] M. S. Ali, H. Tabassum and E. Hossain, "Dynamic User Clustering and Power Allocation for Uplink and Downlink Non-Orthogonal Multiple Access (NOMA) Systems," *IEEE Access,* vol. 4, pp. 6325-6343, 2016.

[69] K. Wang, Y. Liu, Z. Ding, A. Nallanathan, and M. Peng, "User association and power allocation for multi-cell non-orthogonal multiple access networks," *IEEE Transactions on Wireless Communications*, vol. 18, no. 11, pp. 5284-5298, Nov. 2019.

[70] W. Xu, X. Li, C.-H. Lee, M. Pan, and Z. Feng, "Joint sensing duration adaptation, user matching, and power allocation for cognitive OFDM NOMA systems," *IEEE Transactions on Wireless Communications,* vol. 17, no. 2, pp. 1269-1282, Feb. 2018.

[71] D. Fan, F. Gao, G. Wang, Z. Zhong, and A. Nallanathan, "Channel estimation and transmission strategy for hybrid mmWave NOMA systems," IEEE J. Sel. Topics Signal Process., vol. 13, no. 3, pp. 584-596, Jun. 2019.

[72] J. Kim, J. Kim and S. -H. Park, "Joint Design of Power Control and SIC Decoding Order for Max-Min Fairness optimization in Uplink NOMA Systems," *2021 International Conference on Information Networking (ICOIN),* Jeju Island, Korea (South), 2021, pp. 339-342.

[73] 3GPP "Study on Downlink Multiuser Superposition Transmission for LTE," March 2015.

[74] Z. Zhang, H. Sun and R. Q. Hu, "Downlink and Uplink Non-Orthogonal Multiple Access in a Dense Wireless Network," *IEEE Journal on Selected Areas in Communications,* vol. 35, no. 12, pp. 2771-2784, Dec. 2017.

[75] S. Al-Ahmadi, "On the Achievable Max-Min Rates of Cooperative Power-Domain NOMA Systems," *IEEE Access,* vol. 8, pp. 173112-173122, 2020.

[76] Z. Ding, P. Fan and H. V. Poor, "Impact of User Pairing on 5G Non-orthogonal Multiple-Access Downlink Transmissions," *IEEE Transactions on Vehicular Technology,* vol. 65, no. 8, pp. 6010-6023, Aug. 2016.

[77] S. Arzykulov, T. A. Tsiftsis, G. Nauryzbayev and M. Abdallah, "Outage Performance of Cooperative Underlay CR-NOMA With Imperfect CSI," *IEEE Communications Letters,* vol. 23, no. 1, pp. 176-179, Jan. 2019.

[78] L. Dai, B. Wang, Y. Yuan, S. Han, I. Chih-lin and Z. Wang, "Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends," *IEEE Communications Magazine,* vol. 53, no. 9, pp. 74-81, September 2015.

[79] Z. Liu, and L. Yang, "Sparse or dense: a comparative study of code-domain NOMA systems," *arXiv preprint,* arXiv:2009.04148, 2020.

[80] R. Hoshyar, F. P. Wathan and R. Tafazolli, "Novel Low-Density Signature for Synchronous CDMA Systems Over AWGN Channel," *IEEE Transactions on Signal Processing,* vol. 56, no. 4, pp. 1616-1626, April 2008.

[81] H. Nikopour and H. Baligh, "Sparse code multiple access," *IEEE* 24^{th} *Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC),* London, UK, 2013, pp. 332-336.

[82] Z. Yuan, G. Yu, W. Li, Y. Yuan, X. Wang and J. Xu, "Multi-User Shared Access for Internet of Things," *IEEE* 83^{rd} *Vehicular Technology Conference (VTC Spring),* Nanjing, China, 2016, pp. 1-5.

[83] X. Qin and R. Berry, "Opportunistic splitting algorithms for wireless networks," *in Proc. IEEE INFOCOM, Hong Kong, P. R. China,* March 2004, pp. 1662-1672.

[84] M. E. Tanab, W. Hamouda, and Y. Fahmy, "Distributed opportunistic scheduling for MIMO underlay cognitive radio networks," *Journal of Wireless Communications and Mobile Computing,* vol. 16, no. 15, pp. 2212-2224, Oct. 2016.

[85] X. Qin and R. Berry, "Opportunistic Splitting Algorithms for Wireless Networks with Fairness Constraints," 4^{th} *International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks, Boston, MA, USA,* 2006, pp. 1-8.

[86] M. E. Tanab, W. Hamouda and Y. Fahmy, "On the distributed resource allocation of MIMO cognitive radio networks," *IEEE Global Communications Conference (GLOBECOM), San Diego, CA, USA,* Dec. 2015, pp. 1-6.

[87] S. K. Sharma and X. Wang, "Collaborative Distributed Q-Learning for RACH Congestion Minimization in Cellular IoT Networks," *IEEE Communications Letters,* vol. 23, no. 4, pp. 600-603, April 2019.

[88] L. M. Bello, P. D. Mitchell and D. Grace, "Intelligent RACH Access Techniques to Support M2M Traffic in Cellular Networks," *IEEE Transactions on Vehicular Technology,* vol. 67, no. 9, pp. 8905-8918, Sept. 2018.

[89] M. V. da Silva, R. D. Souza, H. Alves and T. Abrão, "A NOMA-Based Q-Learning Random Access Method for Machine Type Communications," *IEEE Wireless Communications Letters,* vol. 9, no. 10, pp. 1720-1724, Oct. 2020.

[90] Z. Chen and D. B. Smith, "Heterogeneous Machine-Type Communications in Cellular Networks: Random Access Optimization by Deep Reinforcement Learning," *IEEE International Conference on Communications (ICC), Kansas City, MO, USA,* 2018, pp. 1-6.

[91] L. Tello-Oquendo, D. Pacheco-Paramo, V. Pla and J. Martinez-Bauset, "Reinforcement Learning-Based ACB in LTE-A Networks for Handling Massive M2M and H2H Communications," *IEEE International Conference on Communications (ICC), Kansas City, MO, USA,* 2018, pp. 1-7.

[92] A. S. A. El-Hameed and K. M. F. Elsayed, "A Q-learning approach for machine-type communication random access in LTE-Advanced," *Telecommunication Systems,* vol. 71, no. 3, pp. 397-413, 2019.

[93] B. Nikfar and A. J. Han Vinck, "Relay selection in cooperative power line communication: A multi-armed bandit approach," *Journal of Communications and Networks,* vol. 19, no. 1, pp. 1-9, February 2017.

[94] F. Li, D. Yu, H. Yang, J. Yu, H. Karl and X. Cheng, "Multi-Armed-Bandit-Based Spectrum Scheduling Algorithms in Wireless Networks: A Survey," *IEEE Wireless Communications,* vol. 27, no. 1, pp. 24-30, February 2020.

[95] A. Ortiz, A. Asadi, M. Engelhardt, A. Klein and M. Hollick, "CBMoS: Combinatorial Bandit Learning for Mode Selection and Resource Allocation in D2D Systems," *IEEE Journal on Selected Areas in Communications,* vol. 37, no. 10, pp. 2225-2238, Oct. 2019

[96] S. Hashima, K. Hatano, E. Takimoto and E. Mahmoud Mohamed, "Neighbor Discovery and Selection in Millimeter Wave D2D Networks Using Stochastic MAB," *IEEE Communications Letters,* vol. 24, no. 8, pp. 1840-1844, Aug. 2020.

[97] J. Ma, T. Nagatsuma, S. Kim and M. Hasegawa, "A Machine-Learning-Based Channel Assignment Algorithm for IoT," *International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Okinawa, Japan,* 2019, pp. 1-6.

[98] M. A. Adjif, O. Habachi and J. -P. Cances, "Joint Channel Selection and Power Control for NOMA: A Multi-Armed Bandit Approach," *IEEE Wireless Communications and Networking Conference Workshop (WCNCW), Marrakech, Morocco,* 2019, pp. 1-6.

[99] P. Zhu, J. Li, D. Wang and X. You, "Machine-Learning-Based Opportunistic Spectrum Access in Cognitive Radio Networks," *IEEE Wireless Communications,* vol. 27, no. 1, pp. 38-44, February 2020.

[100] M. E. Tanab, and W. Hamouda, "Resource Allocation for Underlay Cognitive Radio Networks: A survey," *IEEE Communications Surveys & Tutorials,* vol. 19, no. 2, pp. 1249-1276, Secondquarter 2017.

[101] S. Bubeck and A. Slivkins, "The best of both worlds: stochastic and adversarial bandits," 25^{th} *Annual Conference on Learning Theory (COLT),* 2012.

[102] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-Time Analysis of the Multiarmed Bandit Problem,"*Machine Learning,* vol. 47, no. 2-3, 2002, p. 235–256.

[103] S. Bubeck and N. Cesa-Bianchi, "Regret Analysis of Stochastic and Nonstochastic Multi-Armed Bandit Problems," *Foundations and Trends in Machine Learning,* vol. 5, 2012.

[104] R. Agrawal, "Sample Mean Based Index Policies with O(log N) Regret for the Multi-Armed Bandit Problem." *Advances in Applied Probability,* vol 27, no. 4, pp. 1054-1078, 1995.

[105] M. Vilgelm, S. Schiessl, H. Al-Zubaidy, W. Kellerer and J. Gross, "On the Reliability of LTE Random Access: Performance Bounds for Machine-to-Machine Burst Resolution Time," *IEEE International Conference on Communications (ICC), Kansas City, MO,* May 2018, pp. 1-7.

[106] S. Y. Lien, T. H. Liau, C. Y. Kao, and K. C. Chen, "Cooperative access class barring for machine-to-machine communications," *IEEE Transactions on Wireless Communications,* vol. 11, no. 1, pp. 27-32, Jan. 2012.

[107] S. Duan, V. Shah-Mansouri, Z. Wang and V. W. S. Wong, "D-ACB: Adaptive Congestion Control Algorithm for Bursty M2M Traffic in LTE Networks," *IEEE Transactions on Vehicular Technology,* vol. 65, no. 12, pp. 9847-9861, Dec. 2016.

[108] F. Morvari and A. Ghasemi, "Two-stage resource allocation for random access M2M communications in LTE network," *IEEE Communications Letters,* vol. 20, no. 5, pp. 982-985, May 2016.

[109] W. Zhan and L. Dai, "Massive Random Access of Machine-to-Machine Communications in LTE Networks: Modeling and Throughput Optimization," *IEEE Transactions on Wireless Communications,* vol. 17, no. 4, pp. 2771-2785, April 2018.

[110] K. S. Ko *et al.*, "A novel random access for fixed-location machine-to-machine communications in OFDMA based systems," *IEEE Communications Letters,* vol. 16, no. 9, pp. 1428-1431, Sep. 2012.

[111] J. P. Cheng, C. h. Lee, and T. M. Lin, "Prioritized random access with dynamic access barring for RAN overload in 3GPP LTE-A networks," *2011 IEEE GLOBECOM Workshops (GC Wkshps), Houston, TX,* Dec. 2011, pp. 368-372.

[112] Y. Liu *et al.*, "Design of a scalable hybrid MAC protocol for heterogeneous M2M networks," *IEEE Internet of Things Journal,* vol. 1, no. 1, pp. 99-111, Feb. 2014.

[113] C. Y. Hsu, C. H. Yen, and C. T. Chou, "An adaptive multichannel protocol for large-scale machine-to-machine (M2M) networks," *2013 9th Int. Wireless Commun. and Mobile Comput. Conf. (IWCMC), Sardinia,* July 2013, pp. 1223-1228.

[114] A. K. Gupta and S. Nadarajah, Handbook of Beta Distribution and Its Applications. CRC Press, 2004.

[115] K. E. Atkinson, *An introduction to numerical analysis,* 2nd ed. New York: John Wiley and Sons, 1989.

[116] J. L. Massey, *Collision-resolution algorithms and random-access communications,* Multiuser communication systems. Springer Vienna, 1981.

[117] D. Bertsekas, and R Gallager, *Data networks,* 2nd Ed., Prentice Hall, 1992.

[118] N. Zhang, G. Kang, J. Wang, Y. Guo and F. Labeau, "Resource allocation in a new random access for M2M communications," *IEEE Communications Letters,* vol. 19, no. 5, pp. 843-846, May 2015.

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

[120] D. Johnson, "The triangular distribution as a proxy for the beta distribution in risk analysis," *J. Roy. Statist. Soc. D (Statistician),* vol. 46, no. 3, pp. 387-398, 1997.

[121] S. Kotz, and J. R.Van Dorp, *Beyond Beta: Other Continuous Families of Distributions with Bounded Support and Applications*, World Scientific, 2004.

[122] O. Arouk, A. Ksentini, and T. Taleb, "Performance analysis of RACH procedure with Beta traffic-activated machine-type-communication," *IEEE Global Communications Conference (GLOBECOM), San Diego, CA,* Dec. 2015, pp. 1-6.

[123] D. Niyato, L. Xiao, and P. Wang, "Machine-to-machine communications for home energy management system in smart grid," *IEEE Communications Magazine,* vol. 49, no. 4, pp. 53-59, April 2011.

[124] M. I. Hossain, A. Azari, J. Markendahl and J. Zander, "Enhanced Random Access: Initial access load balance in highly dense LTE-A networks for multiservice (H2H-MTC) traffic," 2017 IEEE International Conference on Communications (ICC), 2017, pp. 1-7.

[125] F. Ghavimi, Y. Lu, and H. Chen, "uplink scheduling and power allocation for M2M communications in SC-FDMA-based LTE-A networks with QoS guarantees", *IEEE Transactions on Vehicular Technology,* vol. 66, no. 7, pp. 6160-6170, July 2017.

[126] A. E. Mostafa, Y. Zhou and V. W. S. Wong, "Connectivity maximization for narrowband IoT systems with NOMA," *IEEE International Conference on Communications (ICC)*, Paris, 2017, pp. 1-6.

[127] M. Shehab, A. K. Hagelskjær, A. E. Kalør, P. Popovski and H. Alves, "Traffic Prediction Based Fast Uplink Grant for Massive IoT," *IEEE* 31^{st} *Annual International Symposium on Personal, Indoor and Mobile Radio Communications,* London, United Kingdom, 2020, pp. 1-6.

[128] S. Ali, A. Ferdowsi, W. Saad and N. Rajatheva, "Sleeping Multi-Armed Bandits for Fast Uplink Grant Allocation in Machine Type Communications,"*IEEE Globecom Workshops (GC Wkshps), Abu Dhabi, United Arab Emirates,* Dec. 2018, pp. 1-6.

[129] M. Shirvanimoghaddam, M. Condoluci, M. Dohler and S. J. Johnson, "On the Fundamental Limits of Random Non-Orthogonal Multiple Access in Cellular Massive IoT,"

IEEE Journal on Selected Areas in Communications, vol. 35, no. 10, pp. 2238-2252, Oct. 2017.

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

[131] Z. Ding, R. Schober, P. Fan and H. V. Poor, "Simple Semi-Grant-Free Transmission Strategies Assisted by Non-Orthogonal Multiple Access," *IEEE Transactions on Communications,* vol. 67, no. 6, pp. 4464-4478, June 2019.

[132] 3GPP, "Radio frequency (RF) requirements for LTE pico node B," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 36.931.

[133] C. Bisdikian, L. M. Kaplan, and M. B. Srivastava, "On the quality and value of information in sensor networks," *ACM Transactions on Sensor Networks (TOSN),* vol. 9, no. 4,pp. 48:1-48:26, Jul. 2013.

[134] P. Auer, "Using confidence bounds for exploitation-exploration trade-offs," *Jrnl. Machine Learning Research,* vol. 3, pp. 397-422, Nov. 2002.

[135] M. A. Sedaghat and R. R. Müller, "On User Pairing in Uplink NOMA," *IEEE Transactions on Wireless Communications,* vol. 17, no. 5, pp. 3474-3486, May 2018.

[136] Q. Zhao and L. Tong, "Opportunistic carrier sensing for energy-efficient information retrieval in sensor networks," *EURASIP Journal on Wireless Communications and Networking,* vol. 2, no. 2, pp. 1-11, April 2005.