Optimized Resource Allocation Techniques for Critical Machine-Type Communications in Mixed LTE Networks

by

©Mohammed Younis Mohammed Abdelsadek

A dissertation submitted to the School of Graduate Studies
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Faculty of Engineering and Applied Science
Memorial University of Newfoundland

May 2020

St. John's Newfoundland

Abstract

To implement the revolutionary Internet of Things (IoT) paradigm, the evolution of the communication networks to incorporate machine-type communications (MTC), in addition to conventional human-type communications (HTC) has become inevitable. Critical MTC, in contrast to massive MTC, represents that type of communications that requires high network availability, ultra-high reliability, very low latency, and high security, to enable what is known as mission-critical IoT. Due to the fact that cellular networks are considered one of the most promising wireless technologies to serve critical MTC, the International Telecommunication Union (ITU) targets critical MTC as a major use case, along with the enhanced mobile broadband (eMBB) and massive MTC, in the design of the upcoming generation of cellular networks. Therefore, the Third Generation Partnership Project (3GPP) is evolving the current Long-Term Evolution (LTE) standard to efficiently serve critical MTC to fulfill the fifth-generation (5G) requirements using the evolved LTE (eLTE) in addition to the new radio (NR). In this regard, 3GPP has introduced several enhancements in the latest releases to support critical MTC in LTE, which is designed mainly for HTC. However, guaranteeing stringent quality-of-service (QoS) for critical MTC while not sacrificing that of conventional HTC is a challenging task from the radio resource management perspective.

In this dissertation, we optimize the resource allocation and scheduling process for critical MTC in mixed LTE networks in different operational and implementation cases. We target maximizing the overall system utility while providing accurate

guarantees for the QoS requirements of critical MTC, through a cross-layer design, and that of HTC as well. For this purpose, we utilize advanced techniques from the queueing theory and mathematical optimization. In addition, we adopt heuristic approaches and matching-based techniques to design computationally-efficient resource allocation schemes to be used in practice. In this regard, we analyze the proposed methods from a practical perspective. Furthermore, we run extensive simulations to evaluate the performance of the proposed techniques, validate the theoretical analysis, and compare the performance with other schemes. The simulation results reveal a close-to-optimal performance for the proposed algorithms while outperforming other techniques from the literature.

To Mom and Dad,

To my siblings,

To my wife and my kids, Abdelrahman and Leen

Acknowledgments

First and foremost, I would like to thank the Almighty God for giving me the strength and means to complete my Ph.D. studies.

Also, I would like to express my sincere thanks and deepest gratitude to my supervisors. Thank you, Dr. Mohamed H. Ahmed, for giving me this opportunity and for all your continuous support, guidance, trust, and inspirational mentorship through my program. Special thanks to Dr. Yasser Gadallah for his invaluable advice, help, and dedication. I would also like to thank Dr. Mohamed Shehata for his helpful feedback and encouragement.

I want to acknowledge the financial support provided by my supervisor Dr. Mohamed H. Ahmed, School of Graduate Studies, and the Natural Sciences and Engineering Research Council of Canada (NSERC).

Last but not least, there are no words enough to thank my family for their support. Your love and prayers, mom and dad, gave me the strength to reach this step in my education. I owe you all the success in my life. I would also like to thank my brother and three sisters for their support. I want to thank my wife for her love, patience, and believing in me. Finally, I thank my kids for bringing hope to my life.

Table of Contents

Abstract **ii**

Acknowledgments **v**

Table of Contents **x**

List of Tables **xi**

List of Figures **xiii**

List of Acronyms **xiv**

Co-authorship Statement **1**

1 Introduction and Overview **2**

1.1 Thesis Motivation 2

1.2 Related Studies of Resource Allocation for Critical MTC in LTE . . . 5

1.2.1 Studies of Resource Allocation for Critical MTC Without HTC 6

1.2.2 Studies of Resource Allocation for Critical MTC Considering HTC Coexistence 7

1.2.3 Studies Consider sTTI or FBC 9

1.3 Research Gap and Problem Statement 11

1.4 Thesis Contributions 12
1.4.1 Key Outcomes 12
1.4.2 List of Publications 14
1.5 Thesis Organization 15

References **16**

2 Resource Allocation with Exact Models in Different Operational Cases **22**

2.1 Abstract 22
2.2 Introduction 23
2.2.1 Related Work 25
2.2.2 Contributions and Outline 26
2.3 System Model and Problem Formulation 27
2.4 Optimal Cross-Layer Resource Allocation 32
2.4.1 Case 1: Wideband CQI Reporting for All Users 33
2.4.2 Case 2: Wideband CQI Reporting for M2M Users 37
2.4.3 Analytical Performance Evaluation of Cases 1 and 2 38
2.4.4 Case 3: The General Case 42
2.5 Practical Implementations and Complexity Analysis 45
2.5.1 Practical Implementations 46
2.5.2 Complexity Analysis 46
2.6 Simulation Results 52
2.7 Conclusions 57

References **57**

3 A Two-Sided Matching Approach for Resource Allocation 62

3.1 Abstract . . . 62

3.2 Introduction . . . 63

3.2.1 Related Work . . . 64

3.2.2 Paper Contributions . . . 66

3.3 System Model and Problem Formulation . . . 69

3.3.1 Effective Bandwidth and Effective Capacity . . . 73

3.3.2 Formulation of the Cross-Layer PDBV Constraint . . . 74

3.4 Matching-Based Scheduling . . . 76

3.4.1 BLP Problem Simplification . . . 76

3.4.2 Two-Sided Matching Formulation . . . 78

3.4.3 Matching-Based Scheduling Algorithm . . . 81

3.5 Analysis of the Matching-Based Scheduling . . . 82

3.5.1 Convergence . . . 82

3.5.2 Stability . . . 83

3.5.3 Computational Complexity . . . 84

3.6 Results . . . 85

3.7 Conclusions . . . 88

References 88

4 Resource Allocation in Massive MIMO LTE 92

4.1 Abstract . . . 92

4.2 Introduction . . . 93

4.2.1 Related Work . . . 95

4.2.2 Paper Contributions and Organization . . . 97

4.3 System Model and Problem Formulation . . . 98

4.3.1 System Model and General Formulation 98
4.3.2 Cross-Layer Design and Formulation 102
4.4 Matching-Based Resource Allocation 106
4.4.1 Formulation of the Instantaneous Resource Allocation Problem 106
4.4.2 Matching Model and Formulation 109
4.4.3 Matching-Based Resource Allocation Algorithm 112
4.5 Analysis of the Proposed Methods 116
4.5.1 Stability 116
4.5.2 Convergence 117
4.5.3 Optimality 118
4.5.4 Computational Complexity 119
4.6 Experimental Results 121
4.7 Conclusions 127

References **128**

5 Resource Allocation in LTE with FBC and sTTIs **133**
5.1 Abstract 133
5.2 Introduction 134
5.2.1 Related Work 136
5.2.2 Paper Contributions and Outline 137
5.3 System Model 138
5.4 Problem Formulation and Analysis 141
5.4.1 Scheduling of the HTC Traffic 141
5.4.2 Scheduling of the cMTC Traffic 144
5.5 Matching-Based Scheduling for HTC Traffic 154
5.5.1 Matching Setup and Algorithm 154

5.5.2 Matching Analysis 156
5.6 Matching-Based Scheduling for the cMTC Traffic 159
5.6.1 Matching Setup and Algorithm 159
5.6.2 Matching Analysis 160
5.7 Simulations Results 163
5.8 Conclusions 168

References **168**

6 Conclusions and Future Work **172**

6.1 Conclusions 172
6.2 Future Work 174

List of Tables

2.1 Frequently Used Symbols and Notations of Chapter 2 28

2.2 Simulation Parameters of Chapter 2 51

3.1 Summary of Notations of Chapter 3 68

3.2 Simulation Parameters of Chapter 3 85

4.1 Frequently Used Symbols and Notations of Chapter 4 99

4.2 Simulation Parameters of Chapter 4 121

5.1 Frequently Used Symbols and Notations of Chapter 5 140

5.2 Simulation Parameters of Chapter 5 163

List of Figures

2.1 Coexistence of critical MTCDs and H2H UEs in a single LTE cell. . . 28

2.2 Comparison of the simulations and analytical results. 52

2.3 Comparison of the proposed algorithms and other scheduling algorithms for Case 1. 53

2.4 Comparison of the proposed algorithms and other scheduling algorithms for Case 2. 54

2.5 Comparison of the proposed algorithms and other scheduling algorithms for Case 3. 55

3.1 An LTE cell with critical MTC coexistent with HTC. 68

3.2 Uplink cross-layer scheduling. 72

3.3 Aggregate HTC achievable data rate. 86

3.4 Average PDBV of the MTCDs. 87

4.1 An eNB with massive antennas serving critical MTCDs coexistent with HTC UEs in an LTE cell. 98

4.2 A cross-layer perspective of the eNB scheduler. 103

4.3 Weighted directed friendship network between users. 110

4.4 Aggregate HTC achievable data rate. 122

4.5 Comparison with the global optimal solution (3 runs and 100 TTIs). . 123

4.6 Average PDBV of MTCDs in the cell. 124
4.7 CDF of the major parameters of the matching algorithm based on 28,315 samples. 127

5.1 HTC and cMTC coexistence in an LTE cell and frame structure. . . . 139
5.2 Queueing model at the eNB. 146
5.3 Performance evaluation and comparison for Case 1. 164
5.4 Performance evaluation and comparison for Case 2. 166

List of Acronyms

AMC Adaptive Modulation and Coding

AWGN Additive Wight Gaussian Noise

BB Branch and Bound

BLP Binary Linear Program

BNLP Binary Non-Linear Program

CDF Cumulative Distribution Function

CMA Cumulative Moving Average

cMTC critical MTC

CQI Channel Quality Indicator

EB Effective Bandwidth

EC Effective Capacity

eLTE evolved LTE

eMBB enhanced Mobile Broadband

eNB evolved Node B

FBC Finite Blocklength Coding

FDMA Frequency-Division Multiple Access

GA Genetic Algorithm

H2H Human-to-Human

ITU International Telecommunication Union

IoT Internet of Things

HTC Human-Type Communications

LTE Long-Term Evolution

MAC Medium Access Control

MIMO Multiple-Input Multiple-Output

MTC Machine-Type Communications

MTCD MTC Device

M2M Machine-to-Machine

NR New Radio

OFDMA Orthogonal Frequency Division Multiple Access

PDBV Probability of Delay-Bound Violation

PDF Probability Density Function

PF Proportional Fairness

PHY PHYsical

PMF Probability Mass Function

PRB Physical Resource Block

PSD Power Spectral Density

PTE Probability of Transmission Error

QoS Quality-of-Service

RRM Radio Resource Management

SNR Signal-To-Noise Ratio

sTTI shortened Transmission Time Interval

TTI Transmission Time Interval

UE User Equipment

URLLC Ultra-Reliable Low-Latency Communications

3GPP Third-Generation Partnership Project

Co-authorship Statement

I, Mohammed Abdelsadek, have the principal authorship status for all the manuscripts included in this dissertation. My supervisors, Dr. Mohamed H. Ahmed and Dr. Yasser Gadallah are the co-authors of all the manuscripts. The list of the manuscripts included in this dissertation is as follows:

- Chapter 2: M. Y. Abdelsadek, Y. Gadallah, and M. H. Ahmed, "Optimal Cross-Layer Resource Allocation for Critical MTC Traffic in Mixed LTE Networks," *in IEEE Transactions on Vehicular Technology,* vol. 68, no. 6, pp. 5944–5956, June 2019.

- Chapter 3: M. Y. Abdelsadek, M. H. Ahmed, and Y. Gadallah, "Cross-Layer Resource Allocation for Critical MTC Coexistent with Human-Type Communications in LTE: A Two-Sided Matching Approach," *submitted to IET Communications,* Sept. 2019.

- Chapter 4: M. Y. Abdelsadek, Y. Gadallah, and M. H. Ahmed, "Matching-Based Resource Allocation for Critical MTC in Massive MIMO LTE Networks," *in IEEE Access,* vol. 7, pp. 127141–127153, Sept. 2019.

- Chapter 5: M. Y. Abdelsadek, Y. Gadallah, and M. H. Ahmed, "A Critical MTC Resource Allocation Approach for LTE Networks with Finite Blocklength Codes," *submitted to IEEE Transactions on Vehicular Technology,* Aug. 2019.

Chapter 1

Introduction and Overview

In this chapter, we first discuss the motivation of this dissertation. Then, we present an overview of the related studies in Section 1.2. The research gap and dissertation contributions are described in Sections 1.3 and 1.4, respectively. Finally, the organization of the dissertation is discussed in Section 1.5.

1.1 Thesis Motivation

The Internet is arguably one of the greatest inventions of humankind. It has given the way to some of the most rapid technological breakthroughs in human history. In this regard, the Internet of Things (IoT) is a revolutionary paradigm that extends the network of networks to almost every "thing" that surrounds humans in their daily life. These *things* include sensors, tags, wearables, home appliances, vehicles, traffic controllers, and industrial systems, to name a few. It is expected that more than 50 billion devices will be connected to the IoT by 2020 [1]. This is a direct result of the convergence of different technologies, such as wireless and mobile networks, embedded systems, and real-time analytics.

As a consequence, the evolution of the communication networks to incorporate

machine-type communications (MTC), in addition to conventional human-type communications (HTC) has become inevitable. In this respect, MTC can be categorized into two major classes, massive MTC and critical MTC [2]. The former represents the connectivity of an enormous number of low-cost, low-power smart devices such as tags, sensors, and wearables. This category emphasizes the challenges of energy-efficiency and the support of a large number of devices in a small geographical area. On the other hand, another primary class of IoT devices requires high network availability, ultra-high reliability, very low latency, and high security. Therefore, the interconnectivity of this type of objects is known as mission-critical IoT. This type of communication is targeted under the critical MTC category. This enables applications such as:

- medical treatment including remote surgeries in e-health services;
- emergency and disaster alarms and responses;
- traffic safety and critical messages in vehicle-to-X connectivity;
- remote control of machinery and critical infrastructure that enables the industrial IoT; and
- reliable remote action with robots and drones.

Furthermore, it unleashes many applications that could emerge if the platform that supports such type of mission-critical applications is well-established.

Among different wireless technologies, cellular networks are considered the most suitable to provide the connectivity of critical MTC devices (MTCDs). This is due to several advantages that include:

- ubiquitous coverage;
- the availability of licensed spectrum that can protect such type of critical communication;

- advanced and flexible radio resource management (RRM) techniques and procedures; and

- enabling device interoperation and mass production of standards-compliant equipment as a standardized technology.

In this regard, critical MTC is targeted as a major usage case in the fifth-generation (5G) cellular networks by the International Telecommunication Union (ITU) under the name ultra-reliable low-latency communications (URLLC) [3]. This is in addition to the other target use cases, namely, enhanced mobile broadband (eMBB) and massive MTC. To fulfill the requirements of the International Mobile Telecommunications 2020 and beyond (IMT-2020), the Third Generation Partnership Project (3GPP) has been developing two components of the radio technology. The first is a novel radio known as the New Radio (NR), which is designed to be deployed in the spectrum above 6 GHz. The second one is the evolved Long-Term Evolution (eLTE) that is targeted to fulfill the 5G requirements while providing backward-compatibility to other pre-5G devices [2, 4]. Toward this end, 3GPP has introduced several enhancements in the PHYsical (PHY) and Medium Access Control (MAC) layers in Releases 14 and 15 to support critical MTC in the LTE standard, that is designed primarily for broadband HTC. As discussed in the 3GPP study [5] and work items [6], [7], and [8], these improvements include:

- support of shortened transmission time intervals (sTTIs) in the level of 0.143 ms compared to the legacy TTIs of 1 ms, based on various lengths of 2-symbol, 4-symbol, and 7-symbol transmissions.

- a fast uplink grant is provided by the evolved Node B (eNB) to the critical MTC devices to overcome the extra signaling delay-overhead and reduce the latency as a result.

- reduced processing time requirements and techniques, as discussed in [7].

The characteristics of critical MTC traffic are different from those of HTC in several aspects, such as the data rate, packet size, latency-tolerance, and the reliability requirements. Although some HTC applications are delay-intolerant, such as video streaming and voice-over-IP (VoIP), the latency and reliability requirements are considered moderate compared to that of critical MTC. In addition, the packet size of critical MTC transmissions is smaller than that of HTC. For instance, 5G radio is required to support URLLC in the level of 0.5 ms of average user plane latency and reliability of 99.999% for 32-byte-long packets [4]. Guaranteeing such stringent quality-of-service (QoS) requirements while not sacrificing those of conventional HTC is considered a challenging task from the RRM perspective.

The resource scheduling and allocation process is at the heart of the RRM procedures in LTE. It is crucial in optimizing the system performance as it affects the users' QoS performance and hence, the overall system utility. In this dissertation, we address the design of the radio resource allocation and scheduling process for critical MTC in mixed LTE networks considering several system challenges, cross-layer design, QoS guarantees, and recently introduced techniques, as we discuss in Section 1.4.

1.2 Related Studies of Resource Allocation for Critical MTC in LTE

In this section, we present an overview of the related resource allocation studies from the literature. First, we discuss the works that do not consider the coexistence of HTC traffic. Then, the studies that consider the coexistence with HTC are investigated in Section 1.2.2. Finally, the studies that take into consideration the finite blocklength

coding (FBC) and/or sTTIs are presented in Section 1.2.3.

1.2.1 Studies of Resource Allocation for Critical MTC Without HTC

There are several studies that focus on the resource allocation and scheduling for critical MTC only without considering the coexistence of the HTC traffic. In [9], the authors consider scheduling the MTCDs in LTE-based systems based on their channels conditions, i.e., SNR, taking into account their maximum allowed delay. They suppose that there are no classes for the devices. The authors in [10] propose resource allocation and access control techniques for OFDMA cellular systems targeting minimizing the overall energy consumption of the MTCDs, including the transmission and circuit energy. The authors in [11–13] extend the work in [14], which allocates fixed access grants for MTC devices periodically over time intervals, by considering random event-driven traffic and statistical QoS metrics. However, they do not consider the channel conditions of the MTC and HTC users. In [15], the authors propose a predictive scheduler to reduce the uplink transmission delay for event-based MTC applications. They exploit the possibility of predicting when devices may need to transmit data in some MTC applications. For example, when a sensor is triggered in a particular region within a wireless sensor network (WSN), other sensors close to this sensor have higher possibility of being triggered as well. The authors in [16] target a balance between throughput and delay requirements of MTCDs by adjusting the percentage of time at which each scheduler is utilized. In [17], the authors allow the MTCDs to report the age of the oldest packet in their buffers to the eNB to enable it to calculate the absolute deadline for each packet request. However, this requires a new MAC control element, and the channel conditions of MTCDs are not taken into consideration. In [18, 19], the authors propose to have the MTCDs generate a

statistical priority report that indicates the uniqueness of the information to be sent. This report can be considered in the scheduling process as a priority metric. In [20], the authors propose to cluster the MTC devices according to their transmission protocols and QoS requirements. Their data rates are then maximized while considering their minimum rate requirements. In [21], the authors target energy-efficient scheduling of MTC to maximize the network lifetime in single-carrier frequency division multiple access (SC-FDMA) systems. They consider resources contiguity constraints imposed by SC-FDMA. In [22], the authors propose two new utility functions for the low-latency communication traffic that are based on the remaining lifetime of the head-of-queue packets of the users. Then, they propose algorithms that are based on dynamic programming to maximize these utility functions considering the wideband channel feedback case. The authors in [23] propose a downlink scheduler for reliable low latency user equipment (UEs). First, they subdivide the UEs into two groups, high and low priority, according to the possibility of satisfying their QoS requirements, in terms of maximum delay and packet error rate. So, they firstly serve the UEs which have QoS requirements that can be satisfied in the scheduling period. However, they consider a special case of channel status feedback, in which a wideband reporting is used to the whole bandwidth.

However, this approach of considering the MTCDs without considering the HTC traffic does not consider the impact of satisfying the stringent requirements of critical MTC on HTC.

1.2.2 Studies of Resource Allocation for Critical MTC Considering HTC Coexistence

Some works consider the splitting of the radio resources before scheduling the users. In [24–26], the authors divide the available physical resource blocks (PRBs) into two

groups one for HTC and another one for MTC devices depending on their demands, then, schedules them separately while considering fairness and maximum tolerable delay in the scheduling of MTC devices.

Moreover, dividing the users into classes before scheduling has been adopted in a number of studies. In [27], the authors split the users into two queues that are scheduled with two different algorithms; one queue contains the HTC users and the delay-sensitive MTCDs and the other queue for other delay-insensitive MTCDs. The scheduling of the first queue is based on a metric that is an evolution of frequency domain proportional fair scheduling algorithm. The second queue is scheduled by dividing the system's timeline into cyclic periods, which are determined by the delay requirements of MTCDs. At the beginning of each period, the MTCDs which exceed their delay threshold are scheduled using the round-robin algorithm, then in the remaining time of the period, PRBs are allocated according to channel conditions of MTCDs. The authors in [28] propose a mixed queue model. They first assign resources to high priority traffic such as voice and video services of HTC users and real-time services of MTCDs. Then, normal probability traffic such as buffer video and data service of HTC users are allocated resources. Finally, non-real services of MTCDs are handled. In [29], the authors assign the resources based on delay requirements by splitting the HTC users and MTCDs into classes based on their remaining time to the maximum delay tolerance. However, their heuristic algorithm does not consider the channel conditions nor statistical delay requirements.

Nevertheless, scheduling of MTCDs and HTC users without jointly considering their channel conditions yields a non-optimal allocation. For example, channel conditions of a particular radio resource can be equally good for both an MTCD and an HTC user. Therefore, it is more efficient to assign this resource to the HTC user that has more data in its buffer as long as the delay requirements of the MTCD are

satisfied. In addition, the impact on HTC can be minimized by simultaneous resource allocation for both types of communications.

On the other hand, a number of studies consider the scheduling of both types of traffics simultaneously. In [30, 31], the authors consider the resource allocation and sharing of MTCDs and HTC users targeting maximizing the sum-rate of the users in LTE-Advanced systems, utilizing the device-to-device (D2D) communications technique. They formulate the problem as an interference-aware bipartite graph to solve it. In [32], the authors minimize the sum of transmit power in the uplink transmission for both types of users while considering an end-to-end delay metric for the MTCDs and minimum rate constraints for the HTC users. However, they consider a deterministic guarantee for the delay, which is less accurate. In [33], the authors discuss some enhancements in network architecture to fulfill service requirements for MTC devices and allocate radio resources maximizing the aggregate utility functions of both the HTC and MTC users. However, the utility function is based on the achievable data rate of the user or device. The authors in [34], maximize the overall bits-per-joule capacity of the HTC and MTC users based on the effective capacity concept. They consider statistical QoS guarantees for all users. Then, they propose a sub-optimal solution utilizing the invasive weed optimization algorithm. However, the rate maximization for the MTC devices that transmit small-size packets negatively impacts that of the HTC users.

1.2.3 Studies Consider sTTI or FBC

There are a few recent studies that consider the FBC in the design of the resource allocation schemes. In [35], the authors maximize the energy efficiency in the downlink of frequency division multiple access (FDMA)-based systems that serve URLLC while considering their end-to-end delay and packet loss requirements. This is achieved by

optimizing the transmit power, bandwidth and the number of active antennas. They adopt the finite blocklength analysis in [36, 37] to approximate the achievable data rates of the users. However, they do not consider OFDMA-based systems such as LTE. In [38], the authors maximize the energy efficiency of URLLC in OFDMA-based radio access systems considering their QoS requirements of packet loss and latency. For this purpose, they optimize the packet dropping, power allocation, and bandwidth allocation policies. Similar studies in [39, 40] consider both of the uplink and downlink directions while allocating the resources. The work in [41] extends that in [35, 38] exploiting the multi-user diversity. However, they consider FDMA-based cellular systems similar to [35]. In [42], the authors formulate the resource allocation problem of the downlink transmissions of URLLC devices such that the weighted sum-throughput is maximized while considering the QoS requirements of the URLLC devices. Then, they employ a successive convex optimization algorithm for a sub-optimal solution for the problem. The authors in [43] consider maximizing the admissible critical MTC load and investigate the minimum required bandwidth. This is achieved by optimizing resource allocation and packet re-transmission schemes. Nevertheless, these studies do not consider the coexistence of other types of communications, such as HTC and how their different QoS will be affected.

Furthermore, the recently introduced shortened transmission intervals technique is investigated in a number of studies. In [44], the authors propose a punctured scheduling scheme, a mechanism for recovering the punctured resources and link adaptation techniques for URLLC that are coexistent with eMBB. In [45], the authors investigate the data rate loss of HTC under puncturing and categorize it into linear, convex, and threshold models. Additionally, they study the corresponding resource allocation problems of critical MTC coexistent with HTC. The authors in [46] propose a risk-sensitive based approach to minimize the impact on the HTC after puncturing

their resources while satisfying reliability constraints for critical MTC. However, the previously discussed works do not consider the FBC of critical MTC traffic.

1.3 Research Gap and Problem Statement

The studies that we discussed in Section 1.2 have investigated the resource allocation and scheduling process of critical MTC in LTE networks from different perspectives. However, there are several research gaps that still exist as follows:

- The resource allocation problem should consider the co-existence of the critical MTC and conventional HTC such that the fulfillment of the stringent QoS requirements of critical MTC do not impact those of HTC traffic and the overall system utility. Therefore, the resource allocation problem should be formulated such that the system utility is maximized while satisfying the different QoS requirements of both types of communications.

- The stringent QoS requirements of critical MTC should be fulfilled with accurate guarantees such that the URLLC applications can be designed based on certain guaranteed QoS parameters. This necessitates a cross-layer design for the resource allocation process such that both the PHY and MAC layers' parameters are jointly optimized.

- The resource allocation and scheduling algorithms should be designed in a manner that considers the practical implementation and real-time operation. This is due to the fact that the scheduling process should be executed periodically in a very short time.

- The recently introduced techniques and enhancements in 3GPP releases should be exploited to design and implement the resource allocation process efficiently.

This includes sTTIs and massive multiple-input multiple-output (MIMO) techniques.

- In addition, the implementation challenges resulting from applying the recent techniques should be tackled. For example, the use of finite blocklength coding due to supporting sTTIs entails several challenges to the resource allocation process.

1.4 Thesis Contributions

In this section, we describe the primary contributions of this dissertation to address the previously discussed research gaps in resource allocation and scheduling techniques for critical MTC in LTE.

1.4.1 Key Outcomes

The major outcomes of this dissertation can be summarized as follows:

- We propose a novel formulation for the joint resource allocation process of critical MTC and HTC as an optimization problem. The system utility is maximized considering the traffic characteristics of both types of communications. This is targeted while fulfilling the different QoS requirements of critical MTC and HTC.

- We formulate the resource allocation and scheduling problem in different operational conditions. Specifically, we consider different cases of channel condition feedback of the users, different scheduling techniques such as puncturing scheduling [44], the uplink and downlink of traffic directions and different scenarios of traffic transmission.

- We exploit the massive MIMO techniques, such as beamforming, to design the resource allocation and scheduling process more efficiently.

- We address the design of the scheduling process in different transmission interval scales, i.e., TTIs and sTTIs, which is a recently proposed technique in LTE. This includes considering the accompanying challenges, such as finite blocklength coding [36].

- We adopt a cross-layer design to consider both the buffer dynamics and the PHY layer parameters of the critical MTC devices. For this purpose, we utilize advanced techniques from the queueing theory, such as $M/D/1$ and $M/G/1$ queues, the effective bandwidth theory [47] and the effective capacity theory [48]. This enables providing accurate guarantees for the satisfaction of the QoS requirements of critical MTC in terms of latency and reliability. For the latter, we consider both sources of packets losses, i.e., queueing delay and transmission errors.

- We analytically evaluate the performance and efficiency of the proposed problem formulation in some of the considered scenarios.

- In all the previously mentioned cases, we propose methods from mathematical optimization and characteristics of random processes, such as stationarity and ergodicity, to simplify the formulated optimization problems. Then, we discuss the tools and algorithms that can be used to calculate the global optimal solution of the optimization problems.

- We propose computationally-efficient algorithms to solve the formulated optimization problems in much lower complexity, compared to that of the optimal solution, to be used as practical resource allocation and scheduling schemes. For

this purpose, we utilize some heuristic algorithms and the matching theory [49].

- To prove the feasibility of implementing the proposed algorithms in practice, we analyze them from a practical perspective. More specifically, we prove the convergence and stability of the proposed methods and analyze their optimality and computational complexity.

- We implement extensive system-level simulations on MATLAB to evaluate the performance of the proposed algorithms, validate the theoretical analysis, and compare the performance with other techniques from the literature and the optimal benchmark. For the latter, we use different open-source and commercial optimization tools, such as BARON [50], IBM's CPLEX, MATLAB's Optimization Toolboxes, and Genetic Algorithms.

1.4.2 List of Publications

This dissertation has resulted in the following publications:

1. M. Y. Abdelsadek, M. H. Ahmed, and Y. Gadallah, "An LTE Matching-Based Scheduling Scheme for Critical-MTC with Shortened Transmission Time Intervals," *submitted to IEEE ICC-2020 Conference*, Oct. 2019.

2. M. Y. Abdelsadek, M. H. Ahmed, and Y. Gadallah, "Cross-Layer Resource Allocation for Critical MTC Coexistent with Human-Type Communications in LTE: A Two-Sided Matching Approach," *submitted to IET Communications,* Sept. 2019.

3. M. Y. Abdelsadek, Y. Gadallah, and M. H. Ahmed, "A Critical MTC Resource Allocation Approach for LTE Networks with Finite Blocklength Codes," *submitted to IEEE Transactions on Vehicular Technology,* Aug. 2019.

4. M. Y. Abdelsadek, Y. Gadallah, and M. H. Ahmed, "Matching-Based Resource Allocation for Critical MTC in Massive MIMO LTE Networks," *in IEEE Access,* vol. 7, pp. 127141–127153, Sept. 2019.

5. M. Y. Abdelsadek, Y. Gadallah, and M. H. Ahmed, "Optimal Cross-Layer Resource Allocation for Critical MTC Traffic in Mixed LTE Networks," *in IEEE Transactions on Vehicular Technology,* vol. 68, no. 6, pp. 5944–5956, June 2019.

6. M. Y. Abdelsadek, Y. Gadallah, and M. H. Ahmed, "An LTE-Based Optimal Resource Allocation Scheme for Delay-Sensitive M2M Deployments Coexistent with H2H Users," *2017 IEEE Conference on Computer Communications (IEEE INFOCOM 2017) Workshops*, May 2017.

1.5 Thesis Organization

The remainder of this dissertation is organized as follows. In Chapter 2, we address the resource allocation for critical MTC utilizing exact-models from queueing theory for a cross-layer design. We consider different operational cases of channel condition feedback. Heuristic algorithms are proposed to solve the formulated problems in a computationally-efficient manner.

In Chapter 3, we adopt a matching approach to solve the formulated resource allocation problem, which exhibits a more efficient performance close to the optimal one. In addition, we utilize the effective bandwidth and effective capacity concepts for an accurate cross-layer design to provide guarantees for the satisfaction of the QoS requirements of critical MTC.

In Chapter 4, we investigate the resource allocation problem in massive MIMO LTE. We formulate the optimization problem exploiting beamforming techniques for

an efficient design. However, due to the potential interference between users sharing the same resources, the matching-based techniques become more complex. We discuss how advanced techniques from matching theory can be utilized to solve the formulated problems in an efficient manner.

In Chapter 5, we consider the downlink scheduling of critical MTC in LTE adopting sTTIs, puncturing scheduling, and finite blocklength coding. We formulate the resource allocation problem in different scenarios and propose the corresponding computationally-efficient algorithms to solve them.

Finally, we conclude the dissertation in Chapter 6 and discuss the possible extensions of this work in the future.

References

[1] H. Sundmaeker, P. Guillemin, P. Friess, and S. Woelfflé, "Vision and challenges for realising the internet of things," *Cluster of European Research Projects on the Internet of Things, European Commision*, 2010.

[2] E. Dahlman, S. Parkvall, and J. Skold, *4G, LTE-advanced Pro and the Road to 5G.* Academic Press, 2016.

[3] ITU-R M.2083, *IMT vision - Framework and overall objectives of the future development of IMT for 2020 and beyond*, Sept. 2015.

[4] 3GPP TR 38.913, *Study on scenarios and requirements for next generation access technologies, technical specification group radio access network*, Oct. 2016.

[5] 3GPP TR 36.881 v14.0.0, *Study on latency reduction techniques for LTE (Release 14)*, June 2016.

[6] 3GPP RP-160667, *Work item on L2 latency reduction techniques for LTE*, March 2016.

[7] 3GPP RP-161299, *Work Item on shortened TTI and processing time for LTE*, June 2016.

[8] 3GPP RP-171489, *Ultra Reliable Low Latency Communication for LTE*, June 2016.

[9] A. S. Lioumpas and A. Alexiou, "Uplink scheduling for machine-to-machine communications in lte-based cellular systems," in *2011 IEEE GLOBECOM Workshops (GC Wkshps).* IEEE, 2011, pp. 353–357.

[10] C. Y. Ho and C.-Y. Huang, "Energy-saving massive access control and resource allocation schemes for m2m communications in ofdma cellular networks," *IEEE Wireless Communications Letters*, vol. 1, no. 3, pp. 209–212, 2012.

[11] A. G. Gotsis, A. S. Lioumpas, and A. Alexiou, "Evolution of packet scheduling for machine-type communications over lte: Algorithmic design and performance analysis," in *2012 IEEE Globecom Workshops*. IEEE, 2012, pp. 1620–1625.

[12] ——, "M2m scheduling over lte: Challenges and new perspectives," *IEEE Vehicular Technology Magazine*, vol. 7, no. 3, pp. 34–39, 2012.

[13] ——, "Analytical modelling and performance evaluation of realistic time-controlled m2m scheduling over lte cellular networks," *Transactions on Emerging Telecommunications Technologies*, vol. 24, no. 4, pp. 378–388, 2013.

[14] 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, 2011.

[15] J. Brown and J. Y. Khan, "Predictive resource allocation in the lte uplink for event based m2m applications," in *2013 IEEE International Conference on Communications Workshops (ICC)*. IEEE, 2013, pp. 95–100.

[16] A. Elhamy and Y. Gadallah, "Bat: A balanced alternating technique for m2m uplink scheduling over lte," in *2015 IEEE 81st Vehicular Technology Conference (VTC Spring)*. IEEE, 2015, pp. 1–6.

[17] N. Afrin, J. Brown, and J. Y. Khan, "A delay sensitive lte uplink packet scheduler for m2m traffic," in *2013 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2013, pp. 941–946.

[18] A. E. Mostafa and Y. Gadallah, "A statistical priority-based scheduling metric for m2m communications in lte networks," *IEEE Access*, vol. 5, pp. 8106–8117, 2017.

[19] ——, "Uniqueness-based resource allocation for m2m communications in narrowband iot networks," in *Vehicular Technology Conference (VTC-Fall), 2017 IEEE 86th*. IEEE, 2017, pp. 1–5.

[20] F. Ghavimi, Y.-W. Lu, and H.-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, 2017.

[21] A. Azari and G. Miao, "Network lifetime maximization for cellular-based m2m networks," *IEEE Access*, vol. 5, pp. 18 927–18 940, 2017.

[22] E. Khorov, A. Krasilov, and A. Malyshev, "Radio resource scheduling for low-latency communications in lte and beyond," in *2017 IEEE/ACM 25th International Symposium on Quality of Service (IWQoS)*. IEEE, 2017, pp. 1–6.

[23] ——, "Reliable low latency communications in lte networks," in *Black Sea Conference on Communications and Networking (BlackSeaCom), 2017 IEEE International*. IEEE, 2017, pp. 1–5.

[24] A. M. Maia, D. Vieira, M. F. de Castro, and Y. Ghamri-Doudane, "A mechanism for uplink packet scheduler in lte network in the context of machine-to-machine communication," in *2014 IEEE Global Communications Conference*. IEEE, 2014, pp. 2776–2782.

[25] A. M. Maia, M. F. de Castro, and D. Vieira, "A dynamic lte uplink packet scheduler for machine-to-machine communication," in *2014 IEEE 25th annual international symposium on personal, indoor, and mobile radio communication (PIMRC)*. IEEE, 2014, pp. 1609–1614.

[26] A. M. Maia, D. Vieira, M. F. de Castro, and Y. Ghamri-Doudane, "A fair qos-aware dynamic lte scheduler for machine-to-machine communication," *Computer Communications*, vol. 89, pp. 75–86, 2016.

[27] S. Zhenqi, Y. Haifeng, C. Xuefen, and L. Hongxia, "Research on uplink scheduling algorithm of massive m2m and h2h services in lte," in *Information and communications technologies (IETICT 2013), IET international conference on*. IET, 2013, pp. 365–369.

[28] J. Ding, A. Roy, and N. Saxena, "Smart m2m uplink scheduling algorithm over lte," *Elektronika ir Elektrotechnika*, vol. 19, no. 10, pp. 138–144, 2013.

[29] M. K. Giluka, N. Rajoria, A. C. Kulkarni, V. Sathya, and B. R. Tamma, "Class based dynamic priority scheduling for uplink to support m2m communications in lte," in *Internet of Things (WF-IoT), 2014 IEEE World Forum on*. IEEE, 2014, pp. 313–317.

[30] S. Hamdoun, A. Rachedi, and Y. Ghamri-Doudane, "Radio resource sharing for mtc in lte-a: An interference-aware bipartite graph approach," in *2015 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2015, pp. 1–7.

[31] ——, "A flexible m2m radio resource sharing scheme in lte networks within an h2h/m2m coexistence scenario," in *2016 IEEE international conference on communications (ICC)*. IEEE, 2016, pp. 1–7.

[32] A. Aijaz and A. H. Aghvami, "On radio resource allocation in lte networks with machine-to-machine communications," in *Vehicular Technology Conference (VTC Spring), 2013 IEEE 77th.* IEEE, 2013, pp. 1–5.

[33] K. Zheng, F. Hu, W. Wang, W. Xiang, and M. Dohler, "Radio resource allocation in lte-advanced cellular networks with m2m communications," *IEEE Communications Magazine*, vol. 50, no. 7, pp. 184–192, 2012.

[34] A. Aijaz, M. Tshangini, M. R. Nakhai, X. Chu, and A.-H. Aghvami, "Energy-efficient uplink resource allocation in lte networks with m2m/h2h co-existence under statistical qos guarantees," *IEEE Transactions on Communications*, vol. 62, no. 7, pp. 2353–2365, 2014.

[35] C. Sun, C. She, and C. Yang, "Energy-efficient resource allocation for ultra-reliable and low-latency communications," in *GLOBECOM 2017-2017 IEEE Global Communications Conference*. IEEE, 2017, pp. 1–6.

[36] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Transactions on Information Theory*, vol. 56, no. 5, p. 2307, 2010.

[37] W. Yang, G. Durisi, T. Koch, and Y. Polyanskiy, "Quasi-static multiple-antenna fading channels at finite blocklength," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 4232–4265, 2014.

[38] C. She, C. Yang, and T. Q. Quek, "Cross-layer optimization for ultra-reliable and low-latency radio access networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 1, pp. 127–141, 2018.

[39] C. Sun, C. She, C. Yang, T. Q. Quek, Y. Li, and B. Vucetic, "Optimizing resource allocation in the short blocklength regime for ultra-reliable and low-latency communications," *IEEE Transactions on Wireless Communications*, vol. 18, no. 1, pp. 402–415, 2019.

[40] C. She, C. Yang, and T. Q. Quek, "Joint uplink and downlink resource configuration for ultra-reliable and low-latency communications," *IEEE Transactions on Communications*, vol. 66, no. 5, pp. 2266–2280, 2018.

[41] C. Sun, C. She, and C. Yang, "Exploiting multi-user diversity for ultra-reliable and low-latency communications," in *Globecom Workshops (GC Wkshps), 2017 IEEE*. IEEE, 2017, pp. 1–6.

[42] W. R. Ghanem, V. Jamali, Y. Sun, and R. Schober, "Resource allocation for multi-user downlink urllc-ofdma systems," *arXiv preprint arXiv:1901.05825*, 2019.

[43] A. Anand and G. de Veciana, "Resource allocation and harq optimization for urllc traffic in 5g wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 11, pp. 2411–2421, 2018.

[44] K. I. Pedersen, G. Pocovi, J. Steiner, and S. R. Khosravirad, "Punctured scheduling for critical low latency data on a shared channel with mobile broadband," in *2017 IEEE 86th Vehicular Technology Conference (VTC-Fall)*. IEEE, 2017, pp. 1–6.

[45] A. Anand, G. De Veciana, and S. Shakkottai, "Joint scheduling of urllc and embb traffic in 5g wireless networks," in *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*. IEEE, 2018, pp. 1970–1978.

[46] M. Alsenwi, N. H. Tran, M. Bennis, A. K. Bairagi, and C. S. Hong, "embb-urllc resource slicing: A risk-sensitive approach," *arXiv preprint arXiv:1902.01648*, 2019.

[47] C.-S. Chang and J. A. Thomas, "Effective bandwidth in high-speed digital networks," *IEEE Journal on Selected areas in Communications*, vol. 13, no. 6, pp. 1091–1100, 1995.

[48] D. Wu and R. Negi, "Effective capacity: a wireless link model for support of quality of service," *IEEE Transactions on wireless communications*, vol. 2, no. 4, pp. 630–643, 2003.

[49] M. David, *Algorithmics Of Matching Under Preferences*. World Scientific, 2013, vol. 2.

[50] M. Tawarmalani and N. V. Sahinidis, "A polyhedral branch-and-cut approach to global optimization," *Mathematical Programming*, vol. 103, no. 2, pp. 225–249, 2005.

Chapter 2

Resource Allocation with Exact Models in Different Operational Cases

2.1 Abstract

Machine-Type Communications (MTC) play an important role in implementing and enabling the Internet of Things (IoT). Long Term Evolution (LTE) is a strong candidate technology for the interconnection of the MTC devices. However, to optimize LTE for MTC purposes, several issues need to be addressed. This is due to the different and diverse Quality of Service (QoS) requirements of MTC compared to those of Human-to-Human (H2H) communications. In particular, critical-MTC pose many challenges to radio resource management in LTE. They have stringent QoS requirements that need to be considered without sacrificing the QoS of the H2H traffic. In this chapter, we formulate the resource allocation optimization problem from a cross-layer design perspective to consider both the QoS requirements of critical-MTC and

those of H2H communications. We propose methods to handle the optimization problem to reduce the computational complexity of the optimal solution. Additionally, the performance of the proposed resource allocation algorithms is evaluated analytically. Moreover, more computationally-efficient algorithms are proposed to practically implement the resource allocation process in several operational cases. Finally, the computational complexity of the proposed algorithms is analysed. The simulations results show the superiority of the proposed algorithms and methods compared to other techniques from the literature.

2.2 Introduction

The Internet of Things (IoT) is a revolutionary paradigm that provides the connectivity and accessibility of smart objects or things that surround humans in their daily life. Such objects include sensors, tags, smart transportation terminals, control systems, health care devices, and automated home appliances. More than 50 billion devices are expected to be connected to the IoT by 2020 [1]. Machine-to-machine (M2M) communication, which is connectivity centric, is the automated communication among devices or machines without human intervention. It plays a vital role in implementing the IoT. The global sales volume of M2M systems is expected to reach $50 billion by 2020 [2].

The characteristics of Machine Type Communications (MTC) are different than those of Human-to-Human (H2H) communications in terms of data rate, delay tolerance, mobility, and frequency of transmission. This results in different and diverse Quality of Service (QoS) requirements. From the point of view of communication technology, MTC can be categorized into massive-MTC and critical-MTC [3]. The former emphasizes how to provide service to a large or massive number of MTC devices

(MTCDs) such as sensors and actuators. The traffic of such devices is characterized by low data rates and delay tolerance. On the other hand, critical-MTC represents the MTC applications that require high reliability and generate delay-intolerant traffic. Some examples of such MTC applications include remote monitoring in e-health services, control of critical infrastructure, critical messages in vehicle-to-X connectivity, and some industrial processes. This type of MTC requires efficient resource allocation schemes that consider their critical requirements without impacting those of H2H communications.

Although there are several technologies to interconnect critical MTCDs, cellular networks are considered one of the most promising technologies. This is owing to its ubiquitous coverage, flexible radio resource management (RRM) techniques, and the availability of licensed spectrum that protects such critical communications and services. In this regard, Ultra-Reliable and Low-Latency Communication (URLLC) is targeted as a major usage case in the next generation cellular networks [4]. To fulfill the requirements of the fifth generation (5G) networks, the Third Generation Partnership Project (3GPP) is developing two components of radio technology. In addition to the New Radio (NR), that is designed to be deployed in the spectrum above 6 GHz, 3GPP is evolving the Long Term Evolution (LTE) to achieve the 5G requirements while providing backward-compatibility to other devices [3]. Therefore, many features and enhancements have been introduced through various LTE releases to enable and optimize critical-MTC in LTE networks [5], [6]. This is due to the fact that LTE is mainly designed for broadband H2H communications. These enhancements include uplink access on Medium Access Control (MAC), reduced processing time, and short transmission time intervals [7].

In this chapter we address the resource allocation problem for critical-MTC that are coexistent with H2H communications in LTE.

2.2.1 Related Work

Several studies investigate the resource allocation problem for critical-MTC in LTE in the form of delay-sensitive users that coexist with regular H2H users. In [8], the authors propose making the MTCDs report the age of the oldest packet in their buffers. This enables the eNB to take into account the absolute deadline of packet requests. Similarly in [9] and [10], the authors propose to make the MTCDs report a statistical priority report that indicates the uniqueness of the information to be sent. This report can be considered in the scheduling process as a priority metric. However, such types of reporting require a modification in the MAC protocol in LTE. The authors in [11] propose a heuristic algorithm that considers the average delay tolerance of the MTCDs in terms of end to end packet delay. Nevertheless, they target minimizing the transmit power of the users and do not consider maximizing the throughput of H2H users, which can be inversely affected while satisfying the QoS requirements of the MTCDs.

The authors in [12] consider the delay requirements of the MTCDs with statistical guarantees. Specifically, they restrict the probability of delay-bound violation of the devices under a threshold while maximizing the bits per joule capacity for both the M2M and H2H users. In spite of that, they use an iterative Invasive Weed Optimization (IWO) algorithm for each dual variable associated with the dual optimization problem which increases the complexity of solving the problem and yields a suboptimal solution. In [13], the authors consider the resource allocation of MTCDs coexistent with H2H User Equipments (UEs) with statistical delay requirements of the MTCDs. Nevertheless, the study does not provide a computationally-efficient implementation for the proposed algorithm and does not consider the general scenario of channel quality feedback of users. In [14], the MTCDs and H2H users are divided into classes based on their deadlines. Despite that, the channel conditions of the devices

are not considered.

In addition, several studies consider the resource allocation problem for MTCDs only without considering the coexistence of H2H communications. In [15], the scheduling of MTCDs is based on their statistical delay requirements without considering the channel conditions. In [16], the authors propose to cluster the M2M devices according to their transmission protocols and QoS requirements. Their data rates are then maximized while considering their minimum rate requirements. In [17–19], the authors investigate the resource allocation problem of URLLC in short transmission time regime. Therefore, they adopt data rates achievable in the case of using finite blocklength codes as analyzed in [20, 21]. In [22], the radio resources are split between the H2H UEs and MTCDs, and then, the MTCDs are scheduled based on metrics that consider fairness and maximum allowed delay. In [23], the authors target energy efficient scheduling of MTC to maximize the network lifetime in single-carrier frequency division multiple access (SC-FDMA) systems. They consider resources contiguity constraints imposed by SC-FDMA. The authors in [24] propose scheduling the MTCDs such that the throughput and delay requirements are balanced. However, separating the resource allocation for H2H UEs and MTCDs, as addressed in the aforementioned studies, does not yield an optimized allocation on all users and reduces the gain of multiuser diversity.

2.2.2 Contributions and Outline

The main contributions of this chapter can be summarized as follows:

- We consider both the critical-MTC and H2H communications while formulating the resource allocation optimization problem in an LTE cell. The objective function to be maximized is the aggregate throughput of the H2H traffic as will be discussed in Section 2.3. However, we express constraints to satisfy the QoS

requirements of the MTCDs and H2H UEs. The buffer dynamics of the MTCDs as well as the Physical (PHY) layer parameters are considered from a cross-layer design perspective to guarantee the satisfaction of their critical requirements.

- We subdivide the resource allocation problem into three different operational cases and propose the methods to handle the resulting optimization problems to get the optimal solution with lower complexity. The performance of the proposed resource allocation algorithms is evaluated analytically. After that, the analytical results are validated by simulations. Furthermore, for practical implementation of the resource allocation process:
  - we propose computationally-efficient algorithms to solve the problems in the three operational cases;
  - we analyse the computational complexity of the proposed algorithms in terms of big-O notation.

The remainder of the chapter is organized as follows. In Section 2.3, we discuss the system model and formulate the primary resource allocation problem. In Section 2.4, we subdivide the problem into three operational scenarios and devise the techniques to handle the resulting optimization problems. The practical implementations of solving the optimization problems as well as the complexity analysis are discussed in Section 2.5. In Section 2.6, the performance of the proposed methods is evaluated using simulations. Finally, we conclude the study in Section 2.7.

2.3 System Model and Problem Formulation

We consider the cross-layer resource allocation problem in the uplink direction of a single LTE cell as shown in Fig. 2.1. The cell serves critical-MTC devices coexist with

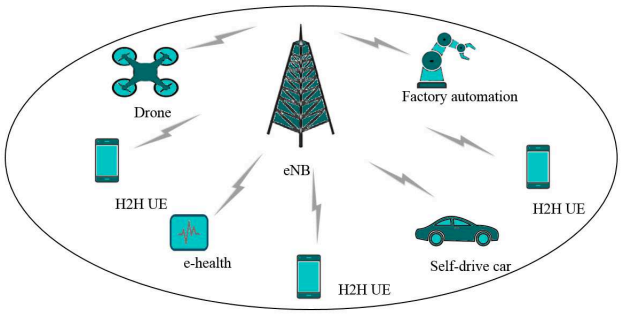


Fig. 2.1: Coexistence of critical MTCDs and H2H UEs in a single LTE cell.

Table 2.1: Frequently Used Symbols and Notations of Chapter 2

Symbol	**Description**
$\mathcal{K}$	Set of PRBs
$\mathcal{K}_u$	Subset of PRBs assigned to user u
$\mathcal{U}$	Set of users
$\mathcal{H}$	Subset of H2H users
$\mathcal{M}$	Subset of MTCDs
K, U, H, M	Cardinalities of $\mathcal{K}, \mathcal{U}, \mathcal{H}, \mathcal{M}$
R_u	Achievable data rate of the uth user
$R_{u,k}$	Achievable data rate of the uth user on kth PRB
R_u^{min}	Required minimum data rate of the uth user
N_u^{max}	Maximum no. of PRBs to be assigned to uth user
$s_{u,k}$	Indicator if kth PRB is assigned to uth user or not
D_u	Delay of uth user
B_u	Delay bound of uth user
V_u	Maximum allowed PDBV of uth user
λ_u	Average arrival rate of uth user
τ_u	Service time of uth user
$R_u^{av,min}$	Required minimum average data rate of uth user
t	TTI index

H2H UEs. Radio resources in LTE are split into Physical Resource Blocks (PRBs). Each PRB has a bandwidth of 180 KHz and can be used for transmission for a period of 1 ms, which is known as the Transmission Time Interval (TTI).

To formulate the resource allocation problem, assume the set of available PRBs at the current TTI is indexed by $\mathcal{K} = \{1, \cdots, k, \cdots, K\}$. This set of PRBs is to be allocated to the users set with index $\mathcal{U} = \{1, \cdots, u, \cdots, U\}$, which is composed of two subsets: a subset $\mathcal{H}$ for H2H UEs, and a subset $\mathcal{M}$ for delay-sensitive MTCDs. The cardinalities of the sets/subsets $\mathcal{K}, \mathcal{U}, \mathcal{H}, \mathcal{M}$ are K, U, H, M, respectively. The frequently used symbols and notations are summarized in Table 2.1.

One important feature of the traffic of critical-MTC is their small packet size [25] and low data rate [26]. However, they have stringent latency requirements that have to be considered in a cross-layer design. That is, after satisfying their latency requirements, their QoS does not improve by increasing their data rate. Nevertheless, the maximization of the data rate of the critical-MTC will be at the cost of that of the H2H traffic. Therefore, the high data rates for such type of critical communications are not important [4], given that their latency requirements are fulfilled. On the other hand, the H2H applications are generally data-hungry, with mainly data rate requirements. That is, their QoS is improved by increasing their data rate. As a consequence, the system utility, that is based on the system throughput, is improved.

Accordingly, we formulate the cross-layer resource allocation problem as a maximization problem. The objective function to be maximized is the aggregate achievable data rate of the H2H UEs. The constraints are expressed such that the QoS requirements of all users are satisfied. That is, the optimization problem can be formulated

as follows:

$$\max_{\{\mathcal{K}_1,\cdots,\mathcal{K}_U\}\in\mathcal{K}} \sum_{u\in\mathcal{H}} R_u \tag{2.1}$$

$$\text{s.t. } \mathcal{K}_u \cap \mathcal{K}_{u'} = \varnothing,\ \forall u \neq u', u, u' \in \mathcal{U} \tag{2.1a}$$

$$R_u \geq R_u^{min},\ \forall u \in \mathcal{H} \tag{2.1b}$$

$$\text{Cross-layer constraint},\ \forall u \in \mathcal{M} \tag{2.1c}$$

$$|\mathcal{K}_u| \leq N_u^{max},\ \forall u \in \mathcal{M}, \tag{2.1d}$$

where R_u is the achievable data rate of user u over the subset of PRBs assigned to it, $\mathcal{K}_u \in \mathcal{K}$. Constraint (2.1a) is used to ensure that every PRB is allocated to only one user. A minimum guaranteed data rate for the H2H UEs is imposed by constraint (2.1b). Constraint (2.1c) is a cross-layer constraint for the QoS requirements of the delay-sensitive MTCDs which will be formulated as discussed in Section 2.4. A maximum allowed number of PRBs that can be assigned to an MTCD is maintained by constraint (2.1d), where $|\mathcal{K}_u|$ is the cardinality of the assigned PRBs subset of the uth user. For instance, in Release 13 of LTE, the maximum number of PRBs that can be allocated to an MTCD is 6 PRBs. It is worth mentioning that non-contiguous resource allocations are allowed in the uplink of LTE-Advanced [27] to enable frequency-selective scheduling in the uplink that increases the spectral efficiency as discussed in [28].

For the mathematical tractability of the problem, we use a $U \times K$ binary indicator matrix $\mathbf{S}$, where $s_{u,k}$ indicates whether the kth PRB is assigned to the uth user.

Therefore, an equivalent optimization problem to (2.1), can be written as follows:

$$\max_{\mathbf{S}} \quad \sum_{u\in\mathcal{H}}\sum_{k=1}^{K} R_{u,k}s_{u,k} \tag{2.2}$$

$$\text{s.t.} \quad \sum_{u=1}^{U} s_{u,k} \leq 1,\ \forall k \in \mathcal{K} \tag{2.2a}$$

$$\sum_{k=1}^{K} R_{u,k}s_{u,k} \geq R_u^{min},\ \forall u \in \mathcal{H} \tag{2.2b}$$

$$\text{Cross-layer constraint},\ \forall u \in \mathcal{M} \tag{2.2c}$$

$$\sum_{k=1}^{K} s_{u,k} \leq N_u^{max},\ \forall u \in \mathcal{M} \tag{2.2d}$$

$$s_{u,k} \in \{0,1\},\ \forall u \in \mathcal{U},\ k \in \mathcal{K}, \tag{2.2e}$$

where $R_{u,k}$ is the achievable data rate of user u over the kth PRB. Constraints (2.2a)–(2.2d) are equivalent to (2.1a)–(2.1d), respectively. Constraint (2.2e) is used to ensure that the decision variable of the optimization problem is binary for all users and PRBs.

To estimate $R_{u,k}$, we first calculate the SNR as follows

$$\gamma_{u,k} = P_{u,k}\frac{|h_{u,k}|^2}{N}, \tag{2.3}$$

where $|h_{u,k}|^2$ is the channel power gain, and $N = N_0 B^{PRB}$ is the noise power calculated from the noise PSD, N_0, and PRB bandwidth, $B^{PRB} = 180$ KHz. $P_{u,k}$ is the transmitted power of the uth user over the kth PRB which is allocated by the resource allocation algorithm. In this chapter, and as in [29], we assume that the transmitted power by every user is equally divided over the assigned PRBs, i.e., $P_{u,k} = \min(P_u/|\mathcal{K}_u|, P_{u,k}^{PRB,max})$, where $P_{u,k}^{PRB,max}$ is the maximum allowed transmit power on the kth PRB by the uth user. After that, the SNR per PRB is quantized and an index number is calculated which is known as Channel Quality Indicator (CQI). According to the LTE standard [30], the CQI reporting can be wideband, where only

one CQI is reported for the entire bandwidth, or at subbands level, which we call here narrowband reporting. Also, in time domain, CQI reporting can be periodic for every TTI, or aperiodic. In the case of discrete data rates as in LTE, where there is a number of supported modulation and coding schemes, the data rate of user u over the kth PRB is given by

$$R_{u,k}(\gamma_{u,k}) = \begin{cases} r_0, & \gamma_{u,k} < \eta_0 \\ r_1, & \eta_0 \leq \gamma_{u,k} < \eta_1 \\ . & . \\ & \\ . & . \\ . & . \\ r_{15}, & \gamma_{u,k} \geq \eta_{15} \end{cases}, \tag{2.4}$$

where $\{\eta_i,\ i = 0, 1, \cdots, 15\}$ are the SNR boundaries for every modulation and coding scheme, and r_i can be calculated using the CQI index values.

2.4 Optimal Cross-Layer Resource Allocation

In this section, we formulate the cross-layer constraint of delay-sensitive MTCDs and present the procedures that can be used to handle the resulting optimization problem in different operational cases of CQI reporting. The general scenario of CQI is to be reported every TTI for every user on every PRB. However, there are other modes of reporting for every user as discussed in Section 2.3. Therefore, we investigate the resource allocation problem for different combinations of CQI modes for critical MTCDs and H2H UEs. In fact, all scenarios can be considered special cases of the general one, and therefore can be addressed using the general formulation. However,

the exact scenario, as a special case, could be exploited to perform the resource allocation process more efficiently. For this purpose, in this chapter, we address the resource allocation problem in two special cases before discussing that in the general scenario. This enables us to formulate the cross-layer constraint of critical-MTC with a more accurate and practical design. In addition, the resulting optimization problems are solved utilizing more computationally-efficient algorithms. Moreover, the performance of the resource allocation is evaluated analytically for these two special cases.

2.4.1 Case 1: Wideband CQI Reporting for All Users

In this case, the CQI reporting is wideband and aperiodic for all users. That is, a single CQI is reported for every user on all the PRBs in the cell for the scheduling period. Therefore, during that scheduling period, the user is serviced with a constant rate that is determined based on the reported wideband CQI. Accordingly, we consider only path-loss and shadowing in the channel model because of the constant channel condition during the scheduling period. Thus, given the channel conditions for every user, the constant service rate in every resource allocation instance is deterministic for the scheduler. According to the 3GPP study in [31], the arrivals of the MTC traffic can be modeled as a Poisson process for triggered M2M applications. Consequently, the buffer dynamics for every MTCD can be modeled as an $M/D/1$ queue which is serviced by a constant rate of R_u. This queuing model, that is similar to that adopted in [32], enables us to formulate the cross-layer constraints for the MTCDs.

As in the framework of LTE QoS [33], statistical QoS guarantees like the probability of delay-bound violation (PDBV) are more accurate and practical than deterministic ones. Therefore, the cross-layer constraint of the MTCDs in this case can be

written as

$$\Pr[D_u > B_u] \leq V_u^{max},\ \forall u \in \mathcal{M}, \tag{2.5}$$

where B_u is the delay-bound, and V_u^{max} is the required maximum probability of delay-bound violation.

To calculate the PDBV in this case, we use the distribution of the waiting time W in the $M/D/1$ queue that is derived in [34] as

$$\Pr[W \leq w] = (1 - \lambda\tau) \sum_{v=0}^{z} \frac{[-\lambda(w - v\tau)]^v}{v!}\ e^{\lambda(w - v\tau)}, \tag{2.6}$$

where τ is the service time, λ is the arrival rate, and z is an integer such that $z\tau \leq w \leq (z+1)\tau$. Therefore, given that $D_u = W_u + \tau_u$, the PDBV of an MTCD, u, can be derived as

$$\Pr[D_u > B_u] = 1 - (1 - \lambda_u\tau_u) \sum_{v=0}^{z} \frac{[-\lambda_u(B_u - \tau_u - v\tau_u)]^v}{v!} e^{\lambda_u(B_u - \tau_u - v\tau_u)}, \tag{2.7}$$

where $\tau_u = 1/R_u$ is the service time of the uth MTCD, and $z\tau_u \leq (B_u - \tau_u) \leq (z+1)\tau_u$.

Using the constraint in (2.5) as the cross-layer constraint in (2.2c), the optimization problem in (2.2) can be reformulated into the following standard form:

$$\max \quad \mathbf{c}^T\mathbf{x} \tag{2.8}$$

$$\text{s.t.} \quad \mathbf{A}\mathbf{x} \leq \mathbf{b} \tag{2.8a}$$

$$\mathbf{g}(\mathbf{x}) \leq \mathbf{V}^{max} \tag{2.8b}$$

$$x \in \{0, 1\}, \tag{2.8c}$$

where the parameters of the optimization problem in (2.8) are related to that of the problem in (2.2) as follows.

The decision variable of the optimization problem in (2.8) is $\mathbf{x} = [\mathbf{s}_1\ \mathbf{s}_2\ \cdots\ \mathbf{s}_U]^T$, where $\mathbf{s}_u$ is the uth row of the indicator matrix $\mathbf{S}$ of the problem in (2.2). The cost vector $\mathbf{c}$ of the standard form in (2.8) can be related to the data rate of users by defining $\mathbf{R}_u = [R_{u,1}\ R_{u,2}\ \cdots\ R_{u,K}]^T$, such that it is given as

$$\mathbf{c} = [\tilde{\mathbf{c}}_1\ \tilde{\mathbf{c}}_2\ \cdots\ \tilde{\mathbf{c}}_U]^T, \tag{2.9}$$

$$\tilde{\mathbf{c}}_u = \begin{cases} \mathbf{R}_u^T, & u \in \mathcal{H} \\ \mathbf{0}_K^T, & \text{otherwise} \end{cases}, \tag{2.10}$$

where $\mathbf{0}_K$ is the K-length zeros vector. This is done to maximize the aggregate data rate of the H2H users only as discussed in Section 2.3.

The matrix of linear inequality constraints of the standard problem in (2.8), $\mathbf{A}$, is composed of three parts as follows

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}^a \\ \mathbf{A}^b \\ \mathbf{A}^c \end{pmatrix}, \tag{2.11}$$

where $\mathbf{A}^a$, $\mathbf{A}^b$, and $\mathbf{A}^c$ represent the left-hand-side of the constraints in (2.2a), (2.2b), and (2.2d), respectively, of the problem in (2.2). The first part, $\mathbf{A}^a = [\mathbf{I}_K\ \mathbf{I}_K\ \cdots\ \mathbf{I}_K]$, is a $K \times KU$ matrix which is composed of U identity matrices of size K. The second part, $\mathbf{A}^b$, is an $H \times KU$ matrix that is composed of the H rows corresponding to the

$\mathcal{H}$ subset of the following matrix

$$\tilde{\mathbf{A}}^{b} = \begin{pmatrix} -\mathbf{R}_1^T & \mathbf{0}_K^T & \cdots & \mathbf{0}_K^T \\ \mathbf{0}_K^T & -\mathbf{R}_2^T & \cdots & \mathbf{0}_K^T \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_K^T & \mathbf{0}_K^T & \cdots & -\mathbf{R}_U^T \end{pmatrix}. \tag{2.12}$$

The last part, $\mathbf{A}^c$, is an $M \times KU$ matrix which is composed of the M rows that correspond to the $\mathcal{M}$ subset of the following matrix

$$\tilde{\mathbf{A}}^{c} = \begin{pmatrix} \mathbf{1}_K^T & \mathbf{0}_K^T & \cdots & \mathbf{0}_K^T \\ \mathbf{0}_K^T & \mathbf{1}_K^T & \cdots & \mathbf{0}_K^T \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_K^T & \mathbf{0}_K^T & \cdots & \mathbf{1}_K^T \end{pmatrix}, \tag{2.13}$$

where $\mathbf{1}_K$ is the K-length ones vector.

The vector $\mathbf{b}$ in the problem in (2.8) has three parts as follows

$$\mathbf{b} = [\mathbf{1}_U^T \quad -\mathbf{R}_{u\in\mathcal{H}}^{min\ T} \quad \mathbf{N}_{u\in\mathcal{M}}^{max\ T}]^T, \tag{2.14}$$

which represent the right-hand-side of the constraints in (2.2a), (2.2b), and (2.2d), respectively, of the problem in (2.2). Finally, constraints vector (2.8b) corresponds to the set of delay constraints (2.2c) defined for MTCDs as in (2.5) and (2.7).

The resulting problem (2.8) falls under the Binary Non-Linear Programs (BNLPs) category. The BNLP is a combinatorial problem and is NP-hard [35]. It can be solved using the Branch and Bound (BB) algorithm which requires high computational complexity and long execution time. However, this instance of a BNLP problem can be efficiently solved using a practical algorithm as discussed in Section 2.5.

2.4.2 Case 2: Wideband CQI Reporting for M2M Users

In this case, we consider the wideband CQI reporting to be done for MTCDs only. All the other assumptions of the last case apply in this case as well. Thus, the modeling of the buffer dynamics of every MTCD as an $M/D/1$ queue is still valid. Also, the cross-layer constraint will be the same, and the resulting BNLP optimization problem as well. However, due to the narrowband CQI reporting of the H2H users in this case, solving the optimization problem requires a more complex algorithm than Case 1. Therefore, we manipulate the optimization problem (2.8) to reduce the computational complexity of the optimal solution.

The high computational complexity of solving problem (2.8) is due to the binary value restriction of the decision variable. Besides that, the relaxations of the problem while applying the BB algorithm are nonlinear which requires complex algorithms to handle the relaxed problems. Thus, we manipulate the optimization problem to get a linear binary program which reduces the computational complexity to a large extent. This can be done by expressing the cross-layer constraints in terms of PHY layer parameters such that we get linear constraints. In other words, we guarantee the required buffer dynamics by restricting the PHY layer parameters, such that the latter can be expressed by linear constraints. That is, constraint (2.8b) is converted into an equivalent rate constraint as follows

$$R_u \geq R_u^{min}(B_u, V_u^{max}),\ \forall u \in \mathcal{M}, \tag{2.15}$$

where $R_u^{min}(B_u, V_u^{max})$ is the minimum data rate that achieves a PDBV that equals the threshold V_u^{max} and can be derived by inverting (2.7) numerically. This numerical inversion is not complex in this case because of the finite discrete values of the service

rate. Consequently, the BNLP is reformulated to be in the following form:

$$\max \quad \mathbf{c}^T\mathbf{x} \tag{2.16}$$

$$\text{s.t.} \quad \tilde{\mathbf{A}}\mathbf{x} \leq \tilde{\mathbf{b}} \tag{2.16a}$$

$$x \in \{0, 1\}, \tag{2.16b}$$

where the constraints matrix $\tilde{\mathbf{A}}$ is composed of three parts; the first and third are as in Case 1, but the second part incorporates minimum rate constraints for all users. That is, the matrix $\tilde{\mathbf{A}}$ can be written as follows

$$\tilde{\mathbf{A}} = \begin{pmatrix} \mathbf{A}^a \\ \tilde{\mathbf{A}}^b \\ \mathbf{A}^c \end{pmatrix}, \tag{2.17}$$

where $\mathbf{A}^a$ and $\mathbf{A}^c$ are the same as in Case 1, and $\tilde{\mathbf{A}}^b$ is given by (2.12). The right hand side vector of the inequality constraints, $\tilde{\mathbf{b}}$, is given by

$$\tilde{\mathbf{b}} = [\mathbf{1}_U^T \quad -\mathbf{R}_{u\in\mathcal{U}}^{min\,T} \quad \mathbf{N}_{u\in\mathcal{M}}^{max\,T}]^T. \tag{2.18}$$

The optimization problem (2.16) is a Binary Linear Program (BLP) that can be solved using the BB algorithm to get the optimal solution but with much lower complexity compared to problem (2.8).

2.4.3 Analytical Performance Evaluation of Cases 1 and 2

In this section, we analytically analyze the performance of the proposed optimal resource allocation techniques discussed in Sections 2.4.1 and 2.4.2. These analytical results are then validated by simulations in Section 2.6. Specifically, we derive the

resulting cell aggregate throughput and the average PDBV for every MTCD in the cell. This is summarized in Theorem 2.4.1.

Theorem 2.4.1. *The optimal solution of the optimization problem given by (2.8) or (2.16), after setting $R_u^{min} \geq \lambda_u, \forall u \in \mathcal{H}$, yields an average cell throughput of*

$$E\{T^{cell}\} = HE\{\lambda_{u\in\mathcal{H}}\} + ME\{\lambda_{u\in\mathcal{M}}\}, \tag{2.19}$$

and an average PDBV for MTCDs approximated by

$$E\{V^{cell}\} \approx \sum_{\forall\lambda_j} \sum_{\forall r_i} g(\lceil \frac{\lambda_j}{r_i} \rceil r_i, \lambda_j) \left[F_{\gamma'}(\eta_i - c) - F_{\gamma'}(\eta_{i-1} - c)\right] f_\lambda[\lambda_j], \tag{2.20}$$

where $\lceil \cdot \rceil$ is the ceil function and $g(R, \lambda)$ is the PDBV as in (2.7), that is a function of the arrival rate λ and service rate R. $f_\lambda[\lambda]$ is the probability mass function (PMF) of the set of arrival rates $\{\lambda_u : u \in \mathcal{M}\}$. $F_{\gamma'}(\gamma')$ is the cumulative distribution function (CDF) of $\gamma' = \gamma - c$, in dB, that is given by

$$F_{\gamma'}(\gamma') = \Phi(q, 0, v) - e^{-q+v^2/2+\ln(\Phi(q,v^2,v))}, \tag{2.21}$$

$$q = \Lambda(\gamma' - \mu), \tag{2.22}$$

$$v = \Lambda\sigma, \tag{2.23}$$

$$\Lambda = \frac{\ln 10}{5n}, \tag{2.24}$$

$$\mu = -10n log_{10} C, \tag{2.25}$$

$$c = P - N - PL0, \tag{2.26}$$

where $\Phi(x, \mu, \sigma)$ is the CDF of a Gaussian RV with mean μ and standard deviation of σ; n is the path loss exponent; $PL0$ is the path loss constant; σ is the standard deviation of shadowing (in dB), and C is the cell radius.

Proof. The proof of (2.19) is straightforward given that the user throughput is calculated as

$$T_u = \begin{cases} R_u, & R_u \leq \lambda_u \\ \lambda_u, & R_u > \lambda_u \end{cases}. \tag{2.27}$$

The optimal solution of the optimization problems (2.8) and (2.16) satisfies the QoS requirements of all users. This implies that $R_u \geq \lambda_u, u \in \mathcal{U}$. Therefore, adding the average throughput of H2H and M2M users yields (2.19) directly.

To prove (2.20), we calculate the expectation of the PDBV for the MTCDs in the cell that have arrival rates of $\{\lambda_u : u \in \mathcal{M}\}$. Given that the PDBV depends on the arrival rate and service rate, that are discrete random variables (RVs), the average PDBV of the MTCDs in the cell can be calculated as follows

$$E\{V^{cell}\} = \sum_{\forall \lambda} \sum_{\forall R} g(R, \lambda) \cdot f_{R,\lambda}[R, \lambda], \tag{2.28}$$

where $g(R, \lambda)$ is the PDBV given by (2.7) which is a function of the arrival rate λ and service rate R. $f_{R,\lambda}[R, \lambda]$ is the joint probability mass function (PMF) of the discrete RVs R and λ of the set of MTCDs. Given that the two RVs R and λ are independent, the PMF $f_{R,\lambda}[R, \lambda]$ can be calculated by multiplying the PMFs of R and λ.

To calculate the PMF of R, we derive the PMF of the data rate per PRB, $f_r[r]$, which originates from the probability density function (PDF) of the SNR per PRB, $f_\gamma(\gamma)$[1]. In Cases 1 and 2, the SNR per PRB, in dB, can be calculated by

$$\gamma = P - N - PL0 - 10n \log_{10} d + x \tag{2.29}$$

$$= c + y + x, \tag{2.30}$$

[1]We drop the subscript of $\gamma_{u,k}$ here to make equations more clear

where $PL0$ is the path loss constant; n is the path loss exponent; x is a Gaussian RV with zero mean and variance σ^2 which represents the shadowing; and $c = P - N - PL0$ is a constant. The RV y is equivalent to $-10n\log_{10} d = -10n\ln d/\ln 10$. Therefore, the cumulative distribution function (CDF) of y is derived as follows

$$F_y(y) = \Pr[d > e^{-y\ln 10/10n}] \tag{2.31}$$

$$= \int_{e^{-y\ln 10/10n}}^{C} \frac{2d}{C^2}\mathrm{d}d \tag{2.32}$$

$$= \frac{1}{C^2}(C^2 - e^{-y\ln 10/5n}), \tag{2.33}$$

where the PDF of d is $2d/C^2$, to yield a uniform user position inside a cell with radius C. Thus, the PDF of y is

$$f_y(y) = \frac{\partial F_y(y)}{\partial y}$$

$$= \frac{1}{C^2}\frac{\ln 10}{5n}e^{-y\ln 10/5n},\ -\frac{10n\ln C}{\ln 10} \leq y < \infty \tag{2.34}$$

Let $\gamma' = \gamma - c$. Therefore, the PDF of γ' is the convolution of the PDFs of the independent RVs y and x. This is mathematically equivalent to the convolution of an exponential RV with rate $\Lambda = \ln 10/5n$ and a Gaussian RV with mean $\mu = -\frac{10n\ln C}{\ln 10} = -10n\log_{10}C$ and variance σ^2 which yields a PDF of an Exponentially Modified Gaussian (EMG) RV given in [36] as

$$f_{\gamma'}(\gamma') = \frac{\Lambda}{2}e^{\frac{\Lambda}{2}(2\mu+\Lambda\sigma^2-2\gamma')}\mathrm{erfc}\left(\frac{\mu+\Lambda\sigma^2-\gamma'}{\sqrt{2}\sigma}\right). \tag{2.35}$$

Accordingly, to find the PMF of the data rate per PRB, $f_r[r]$, we calculate the probability that the SNR falls inside a certain range of SNR levels as in (2.4). This

can be calculated using the CDF of γ' as follows

$$\begin{aligned} f_r[r_i] &= \Pr[\eta_{i-1} < \gamma \leq \eta_i] \\ &= F_{\gamma'}(\eta_i - c) - F_{\gamma'}(\eta_{i-1} - c), \end{aligned} \tag{2.36}$$

where $F_{\gamma'}(\gamma')$ is the CDF of γ', which is an EMG RV, that can be calculated as in [36] as given in (2.21).

Consequently, the average PDBV per MTCD in the cell can be approximated by (2.20).

□

2.4.4 Case 3: The General Case

In this general case, all users report an individual CQI for every PRB. In addition to shadowing, channel fading is also considered. The CQIs are reported every TTI for all users.

To formulate the cross-layer constraint for the MTCDs, we model the buffer dynamics of every MTCD in this case as an $M/G/1$ queue. Although the PDBV delay constraint is more accurate, it is computationally inefficient to be considered in this case. This is due to the complexity of the calculation of the waiting time distribution in this queuing model [37] which is computationally-inefficient to be used in a real-time process such as scheduling. Also, in this chapter we target considering exact models and formulas for the delay of critical MTCDs. Therefore, we do not adopt frameworks such as effective bandwidth [38] that uses large deviations theory to approximate the PDBV of the queues. Accordingly, we use average delay as a cross-layer constraint as

follows

$$E\{D_u\} \leq B_u,\ \forall u \in \mathcal{M}. \tag{2.37}$$

From the analysis of $M/G/1$ queue [39], and assuming that there are available PRBs for MTCDs every TTI, the average delay for uth MTCD, $E\{D_u\}$, is calculated by

$$E\{D_u\} = \frac{\lambda_u E\{\tau_u^2\}}{2(1-\lambda_u E\{\tau_u\})} + E\{\tau_u\},\ u \in \mathcal{M}. \tag{2.38}$$

The resulting resource allocation optimization problem at the tth TTI, with the cross-layer constraint as defined in (2.37) and (2.38), is as follows:

$$\max_{\mathbf{S}} \quad \sum_{u\in\mathcal{H}} \sum_{k=1}^{K} R_{u,k}(t) s_{u,k} \tag{2.39}$$

$$\text{s.t.} \quad \sum_{u=1}^{U} s_{u,k} \leq 1,\ \forall k \in \mathcal{K} \tag{2.39a}$$

$$E\{R_u\} \geq R_u^{av,min},\ \forall u \in \mathcal{H} \tag{2.39b}$$

$$\frac{\lambda_u E\{\tau_u^2\}}{2(1-\lambda_u E\{\tau_u\})} + E\{\tau_u\} \leq B_u,\ \forall u \in \mathcal{M} \tag{2.39c}$$

$$\sum_{k=1}^{K} s_{u,k} \leq N_u^{max},\ \forall u \in \mathcal{M} \tag{2.39d}$$

$$s_{u,k} \in \{0,1\},\ \forall u \in \mathcal{U},\ k \in \mathcal{K}, \tag{2.39e}$$

where $R_u^{av,min},\ u \in \mathcal{H}$ is the minimum average data rate requirement of uth H2H UE.

In the same manner as done in Case 2, we reformulate the BNLP problem in (2.39) to be a BLP problem. This can be achieved by expressing the constraints in terms of PHY layer parameters. That is, we replace the constraints (2.39b) and (2.39c) by constraint (2.40), as in Lemma 2.4.2.

Lemma 2.4.2. *At the tth TTI, constraints (2.39b) and (2.39c) can be replaced by the*

following equivalent single constraint

$$R_u(t) \geq R_u^{min}(t),\ \forall u \in \mathcal{U}, \tag{2.40}$$

where

$$R_u^{min}(t) = tR_u^{av,min} - (t-1)R_u^{av}(t-1), \tag{2.41}$$

$$R_u^{av,min} = \frac{1}{1/\lambda_u + B_u - \sqrt{(1/\lambda_u)^2 + B^2}},\ \forall u \in \mathcal{M} \tag{2.42}$$

$$R_u^{av}(t) = \begin{cases} \frac{R_u(t)+(t-1)R_u^{av}(t-1)}{t}, & t \geq 2 \\ R_u(t), & t = 1 \end{cases}. \tag{2.43}$$

Proof. To prove (2.40), we first convert the average delay constraint (2.39c) into an equivalent average data rate constraint as in constraint (2.39b). This is done by solving the inequality of (2.39c) for $E\{R_u\}$. Substituting $E\{\tau_u^2\} = Var[\tau_u] + (E\{\tau_u\})^2$ and rearranging gives

$$\lambda_u(Var[\tau_u] + (E\{\tau_u\})^2) + 2E\{\tau_u\}(1 - \lambda_u E\{\tau_u\}) - 2B_u(1 - \lambda_u E\{\tau_u\}) \leq 0. \tag{2.44}$$

Solving for $E\{\tau_u\}$ yields

$$E\{\tau_u\} \leq \frac{1}{2}(\frac{2}{\lambda_u} + 2B_u) - \frac{1}{2}\sqrt{(\frac{2}{\lambda_u} + 2B_u)^2 - \frac{8B_u}{\lambda_u} + 4Var[\tau_u]} \tag{2.45}$$

$$= \frac{1}{\lambda_u} + B_u - \sqrt{(\frac{1}{\lambda_u})^2 + B_u^2 + Var[\tau_u]} \tag{2.46}$$

$$\leq \frac{1}{\lambda_u} + B_u - \sqrt{(\frac{1}{\lambda_u})^2 + B_u^2}. \tag{2.47}$$

This inequality can be rearranged to be in the following form

$$E\{R_u\} \geq R_u^{av,min}, \tag{2.48}$$

where $R_u^{av,min}$ is the equivalent minimum average rate requirement for the uth MTCD and is given by (2.42).

After expressing constraint (2.39c) as an equivalent average data rate constraint, we convert all average data rate constraints into equivalent instantaneous data rate constraints. This can be achieved by calculating the average achieved data rate at every TTI using the cumulative moving average (CMA). That is, the average data rate requirement is satisfied every TTI by

$$R_u^{av}(t) \geq R_u^{av,min},\ u \in \mathcal{U}, \tag{2.49}$$

where $R_u^{av}(t)$ is the CMA data rate achieved by the uth user at the tth TTI which can be calculated by (2.43).

Rearranging (2.49) yields the equivalent instantaneous data rate requirement given by (2.40). □

2.5 Practical Implementations and Complexity Analysis

The reformulation of the BNLP problems as BLP problems significantly reduces the complexity of the optimization problem. However, more computationally efficient algorithms are required for the scheduling process. This is because, in LTE, the

execution time of the scheduling algorithm is restricted to be within 1 ms. Therefore, in this section, we propose a more practical implementation for the resource allocation methods discussed in the previous section. After that, the computational complexity of the proposed algorithms is investigated.

2.5.1 Practical Implementations

Due to the wideband CQI reporting of the users in Case 1, the optimization problem (2.8) can be optimally solved with a more computationally efficient method. In Algorithm 2.1, we deal with problem (2.8) as an optimization problem while the number of PRBs assigned to every user, is the decision variable. This solves the problem optimally with reduced computational complexity.

To solve the optimization problem (2.16) of Case 2, Algorithm 2.1 is not valid. This is because the CQI reporting of H2H UEs is no longer wideband. Therefore, in Algorithm 2.2, we consider this narrowband CQI reporting. The complexity is much less than that of the optimal solution. On the other hand, this complexity reduction is achieved at the price of optimality. Algorithm 2.2 is suboptimal in this case. However, it exhibits an optimal performance when $R_u^{min} = \lambda_u,\ u \in \mathcal{H}$, as discussed in Section 2.6.

Algorithm 2.3 is a suboptimal algorithm that solves the optimization problem (2.39) of Case 3 in a more efficient way. This algorithm is based on converting the constraints of average delay and average data rate in the problem into equivalent instantaneous data rate constraints as in Lemma 2.4.2.

2.5.2 Complexity Analysis

The computational complexity of the optimal solution using the BB algorithm depends on the number of the explored nodes. Therefore, the worst case of this complexity

Algorithm 2.1 Solving problem (2.8)

Step 1: Calculate minimum PRB requirements of $\mathcal{H}$ users

for all $u \in \mathcal{H}$ **do**
  $\text{PRB}_u^{min} \leftarrow \lceil R_u^{min}/r_u \rceil$
end for

Step 2: Calculate minimum PRB requirements of $\mathcal{M}$ users

for all $u \in \mathcal{M}$ **do**
  for $q = \lceil \lambda_u/r_u \rceil : N_u^{max}$ **do**
    if $\Pr[D_u > B_u | R_u = qr_u] \leq V_u^{max}$ **or** $q = N_u^{max}$ **then**
      $\text{PRB}_u^{min} \leftarrow q$
      Break this for loop.
    end if
  end for
end for

Step 3: Feasibility test

$\text{PRB}_{tot}^{min} \leftarrow \sum_{u=1}^{U} \text{PRB}_u^{min}$
if $\text{PRB}_{tot}^{min} > K$ **then**
  Problem is infeasible. **Stop**
end if

Step 4: PRBs allocation phase 1

$\mathcal{K}^{P1} \leftarrow \mathcal{K},\ \mathcal{U}^{P1} \leftarrow \mathcal{U}$
for all $u \in \mathcal{U}^{P1}$ **do**
  Allocate PRBs from $\mathcal{K}^{P1}$ such that PRB_u^{min} is satisfied.
  Remove the allocated PRBs from $\mathcal{K}^{P1}$.
end for

Step 5: PRBs allocation phase 2

$\mathcal{K}^{P2} \leftarrow \mathcal{K}^{P1},\ \mathcal{U}^{P2} \leftarrow \mathcal{H}$
for all $k \in \mathcal{K}^{P2}$ **do**
  Allocate PRB k to user $u^* \in \mathcal{U}^{P2}$ such that:
  $u^* = \arg\max \sum_{u \in \mathcal{U}^{P2}} R_u$
end for

can reach that of exhaustive search. To analyze the computational complexity of the proposed algorithms, we calculate the worst case complexity of every step using the big-O notation. After that, the total complexity of the algorithm is determined by the dominating terms.

The worst case computational complexity of every step of Algorithm 2.1 is as follows:

Algorithm 2.2 Solving problem (2.16)

Step 1: PRBs allocation phase 1 (satisfy $\mathcal{H}$ requirements)
$\mathcal{K}^{P1} \leftarrow \mathcal{K},\ \mathcal{U}^{P1} \leftarrow \mathcal{H}$
for all $k \in \mathcal{K}^{P1}$ **do**
  Allocate PRB k to user $u^* \in \mathcal{U}^{P1}$ such that:
  $u^* = \arg\max_{u \in \mathcal{U}^{P1}} R_{u,k}$
  Remove k from $\mathcal{K}^{P1}$.
  if $R_{u^*} \geq R_{u^*}^{min}$ **then**
    Remove u^* from $\mathcal{U}^{P1}$.
  end if
  if $\mathcal{U}^{P1} = \emptyset$ **then**
    Break the for loop.
  end if
end for
Step 2: Feasibility test 1
if $\mathcal{K}^{P1} = \emptyset$ **then**
  Problem is infeasible. **Stop**
end if
Step 3: PRBs allocation phase 2 (maximize data rate of $\mathcal{H}$)
$\mathcal{K}^{P2} \leftarrow \mathcal{K}^{P1},\ \mathcal{U}^{P2} \leftarrow \mathcal{H}$
for all $k \in \mathcal{K}^{P2}$ **do**
  Allocate PRB k to user $u^* \in \mathcal{U}^{P2}$ such that:
  $u^* = \arg\max_{u \in \mathcal{U}^{P2}} R_{u,k}$
end for
Step 4: Determine appropriate PRBs for $\mathcal{M}$ users
$\mathcal{K}^{excess} \leftarrow \emptyset$
for all $u \in \mathcal{H}$ **do**
  Sort the allocated PRBs in Phase 1 and 2 in a descending order according to the achievable data rate on them by the user.
  Determine the minimum PRBs subset that satisfies the minimum data rate requirement and add the remainder PRBs to the $\mathcal{K}^{excess}$ subset.
end for
Sort the PRBs in $\mathcal{K}^{excess}$ in an ascending order according to the achievable data rate on them by the $\mathcal{H}$ users that they are allocated to in Phase 1 and 2.
Step 5: Calculate minimum PRB requirements of $\mathcal{M}$ users
for all $u \in \mathcal{M}$ **do**
  for $q = \lceil \lambda_u / r_u \rceil : N_u^{max}$ **do**
    if $\Pr[D_u > B_u | R_u = q r_u] \leq V_u^{max}$ **or** $q = N_u^{max}$ **then**
      $\mathrm{PRB}_u^{min} \leftarrow q$
      Break the inner for loop.
    end if
  end for
end for
Step 6: Feasibility test 2
$\mathrm{PRB}_{\mathcal{M}}^{min} \leftarrow \sum_{u \in \mathcal{M}} \mathrm{PRB}_u^{min}$
if $\mathrm{PRB}_{\mathcal{M}}^{min} > |\mathcal{K}^{excess}|$ **then**
  Problem is infeasible. **Stop**
end if
Step 7: PRBs allocation phase 3 (satisfy $\mathcal{M}$ requirements)
for all $u \in \mathcal{M}$ **do**
  Allocate PRBs from $\mathcal{K}^{excess}$ subset from the beginning such that PRB_u^{min} is satisfied. Remove the allocated PRBs from the subsets of allocated PRBs of $\mathcal{H}$ users.
end for

Algorithm 2.3 Solving problem (2.39)

Step 1: Calculate $R_u^{min}(t), u \in \mathcal{U}$

Use equations (2.40)–(2.43)

Step 2: PRBs allocation phase 1 (satisfy $\mathcal{H}$ *requirements)*

$\mathcal{K}^{P1} \leftarrow \mathcal{K},\ \mathcal{U}^{P1} \leftarrow \mathcal{H}(R_u^{min}(t) > 0)$

for all $k \in \mathcal{K}^{P1}$ **do**

  Allocate PRB k to user $u^* \in \mathcal{U}^{P1}$ such that:

  $u^* = \arg\max_{u \in \mathcal{U}^{P1}} R_{u,k}(t)$

  Remove k from $\mathcal{K}^{P1}$.

  if $R_{u^*}(t) \geq R_{u^*}^{min}(t)$ **then**

    Remove u^* from $\mathcal{U}^{P1}$.

  end if

  if $\mathcal{U}^{P1} = \emptyset$ **then**

    Break the for loop.

  end if

end for

Step 3: Feasibility test 1

if $\mathcal{K}^{P1} = \emptyset$ **or** $\mathcal{U}^{P1} \neq \emptyset$ **then**

  Problem is infeasible. **Stop**

end if

Step 4: PRBs allocation phase 2 (maximize data rate of $\mathcal{H}$*)*

$\mathcal{K}^{P2} \leftarrow \mathcal{K}^{P1},\ \mathcal{U}^{P2} \leftarrow \mathcal{H}$

for all $k \in \mathcal{K}^{P2}$ **do**

  Allocate PRB k to user $u^* \in \mathcal{U}^{P2}$ such that:

  $u^* = \arg\max_{u \in \mathcal{U}^{P2}} R_{u,k}(t)$

end for

Step 5: Determine appropriate PRBs for $\mathcal{M}$ *users*

$\mathcal{K}^{excess} \leftarrow \emptyset$

for all $u \in \mathcal{H}$ **do**

  Sort the allocated PRBs in Phase 1 and 2 in a descending order according to the achievable data rate on them by the user.

  Determine the minimum PRBs subset that satisfy the minimum data rate requirement and add the remainder PRBs to the $\mathcal{K}^{excess}$ subset.

end for

Sort the PRBs in $\mathcal{K}^{excess}$ in an ascending order according to the achievable data rate on them by the $\mathcal{H}$ users that they are allocated to in Phase 1 and 2.

Step 6: PRBs allocation phase 3 (satisfy $\mathcal{M}$ *requirements)*

$\mathcal{K}^{P3} \leftarrow \mathcal{K}^{excess},\ \mathcal{U}^{P3} \leftarrow \mathcal{M}$

for all $k \in \mathcal{K}^{P3}$ **do**

  Clear the allocation of PRB k.

  Allocate PRB k to $u^* \in \mathcal{U}^{P3}$ such that:

  $u^* = \arg\max_{u \in \mathcal{U}^{P3}} R_{u,k}(t)$

  Remove k from $\mathcal{K}^{P3}$

  if $R_{u^*}(t) \geq R_{u^*}^{min}(t)$ **then**

    Remove u^* from $\mathcal{U}^{P3}$

  end if

  if $\mathcal{U}^{P3} = \emptyset$ **then**

    Break the for loop

  end if

end for

Step 7: Feasibility test 2

if $\mathcal{U}^{P3} \neq \emptyset$ **then**

  Problem is infeasible. **Stop**

end if

- step 1 requires $\mathcal{O}(H)$,
- step 2 requires $\mathcal{O}(MN^{max})$,
- steps 3–4 requires $\mathcal{O}(U)$, and
- step 5 requires $\mathcal{O}(H(K-U))$.

Therefore, the computational complexity of Algorithm 2.1 is linear with the number of users and PRBs.

The computational complexity of the steps of Algorithm 2.2 is:

- steps 1–2 require $\mathcal{O}(H^2)$,
- step 3 requires $\mathcal{O}(H(K-H))$,
- step 4 requires $\mathcal{O}(K)$,
- step 5 requires $\mathcal{O}(MN^{max})$, and
- steps 6–7 require $\mathcal{O}(M)$.

Thus, the total computational complexity of the algorithm is $\mathcal{O}(H^2)$.

The worst case computational complexity of Algorithm 2.3 can be derived as follows:

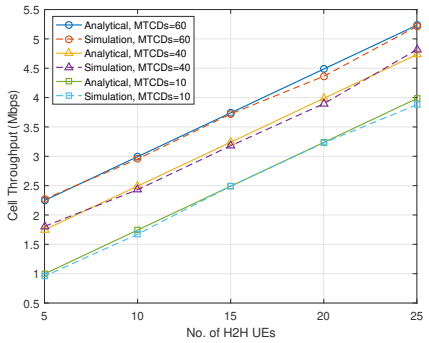
- step 1 requires $\mathcal{O}(U)$,
- steps 2–3 require $\mathcal{O}(H^2)$,
- step 4 requires $\mathcal{O}(H(K-H))$,
- step 5 requires $\mathcal{O}(K)$, and
- steps 6–7 require $\mathcal{O}(M^2)$.

Table 2.2: Simulation Parameters of Chapter 2

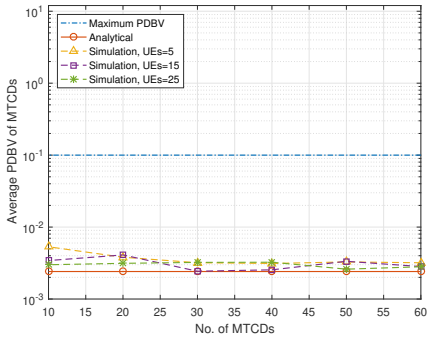
Parameter	**Value**
Cell radius (C)	500 m
Number of eNBs	1
Simulation time	1000 TTIs
Number of runs	30
Path loss ($PL0 + 10n \log_{10} d$)	$128.1 + 37.6 \log_{10}(d)$, d in km [40]
Standard deviation of shadowing (σ)	8 dB [12]
Transmitter power (P)	15 dBm
Power spectral density of noise	-174 dBm/Hz [12]
Noise figure	18 dB
Bandwidth	20 MHz
Number of PRBs (K)	100
Distribution of MTCDs/UEs	Fixed and uniform
H2H arrival rate ($\lambda_u,\ u \in \mathcal{H}$)	$64, 128, 256$ kbps
M2M arrival rate ($\lambda_u,\ u \in \mathcal{M}$)	$10, 20, 30, 40$ kbps [15], [22]
Arrival rates distribution	Uniform
Delay bound ($B_u,\ u \in \mathcal{M}$)	0.2 ms,
Maximum PDBV ($V_u^{max},\ u \in \mathcal{M}$)	10%

Therefore, the total complexity of Algorithm 2.3 is $\mathcal{O}(H^2) + \mathcal{O}(M^2)$.

From the previous analysis, it is apparent that the computational complexity of the proposed algorithms is much lower than that of the optimal solutions. Furthermore, as discussed in Section 2.6, they exhibit the same optimal performance, such as Algorithm 2.1, or close to that of the optimal solution. Therefore, these algorithms represent reasonable practical implementations for the optimal methods discussed in Section 2.4.



(a) Cell throughput versus number of H2H UEs.



(b) Average PDBV of MTCDs versus number of MTCDs.

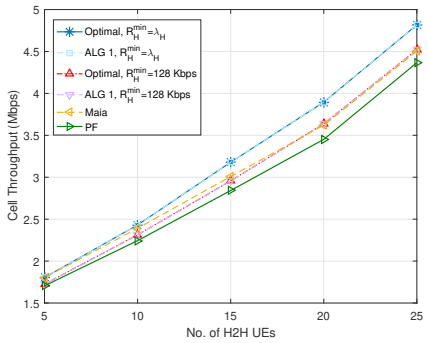
Fig. 2.2: Comparison of the simulations and analytical results.

2.6 Simulation Results

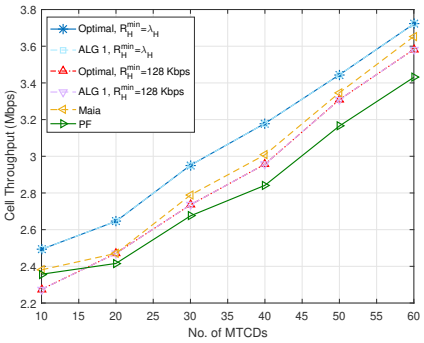
In this section, we present and discuss the simulations results of evaluating the performance of the proposed methods and algorithms. The performance of the proposed resource allocation algorithms is compared with the optimal solution, the proportional fair (PF) scheduler [41] and the resource allocation based on the metrics as in [22] (the first algorithm), referred to as Maia algorithm. For Maia algorithm, we use the same parameters that were used in [22] with the PF scheduler for the H2H users. Also, we validate the analytical results derived in Section 2.4.3 by comparing the values of the derived metrics with that calculated in the simulations.

We perform three different simulation experiments for the three considered cases. In all simulations, we consider the uplink resource allocation of a single LTE cell that includes both critical-MTC and H2H communications. The sources of H2H traffic are Voice-over-IP (VoIP) and video applications. The traffic of critical-MTC is event-based with Poisson arrivals. The arrival rates of the users are uniformly distributed with the values as in Table 2.2 along with other simulation parameters.

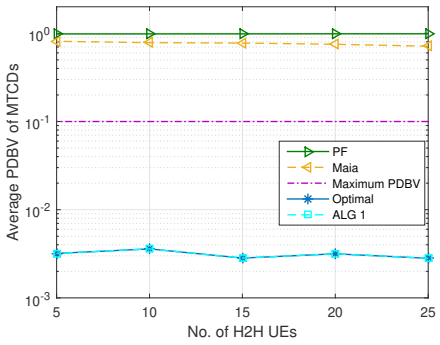
Figure 2.2 shows the cell throughput and average PDBV per MTCD in the cell



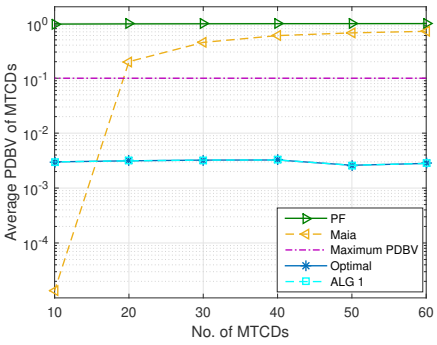
(a) Cell throughput vs. number of H2H UEs ($M = 40$).



(b) Cell throughput vs. number of MTCDs ($H = 15$).



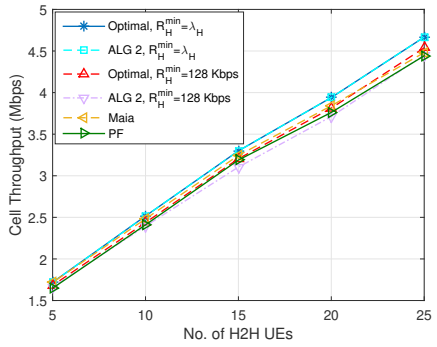
(c) Average PDBV of MTCDs vs. number of H2H UEs ($M = 60$).



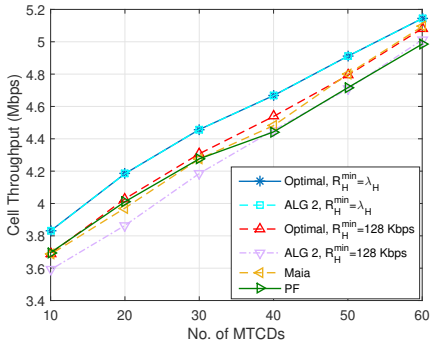
(d) Average PDBV of MTCDs vs. number of MTCDs ($H = 25$).

Fig. 2.3: Comparison of the proposed algorithms and other scheduling algorithms for Case 1.

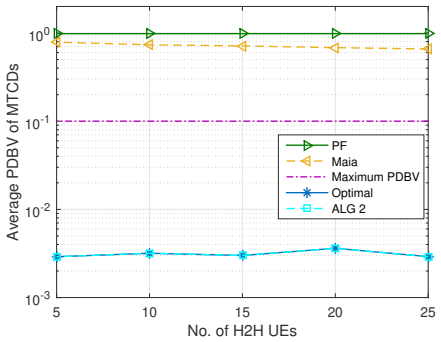
calculated analytically and using simulation. In Fig. 2.2(a), the cell throughput is plotted against the number of H2H UEs at different numbers of MTCDs. The figure shows a close agreement between the analytical and simulation results. In Fig. 2.2(b), the cell average PDBV per MTCD is plotted against the number of MTCDs at different numbers of H2H UEs. Again, the analytical results are quite close to the simulation ones with an acceptable error that decreases with the increase of the number of users. This is apparent by comparing the difference between the results in the case of 60 MTCDs versus the 10 MTCDs case. This agreement in the results



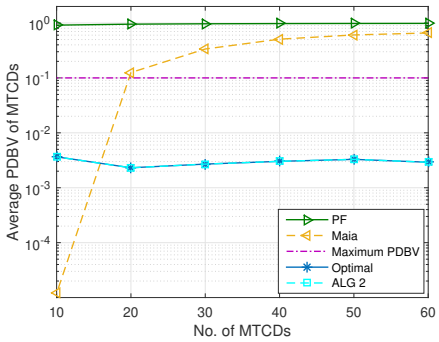
(a) Cell throughput vs. number of H2H UEs ($M = 40$).



(b) Cell throughput vs. number of MTCDs ($H = 25$).



(c) Average PDBV of MTCDs vs. number of H2H UEs ($M = 60$).

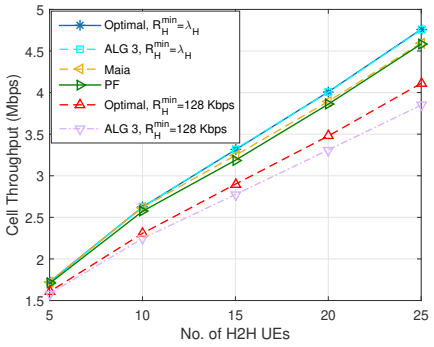


(d) Average PDBV of MTCDs vs. number of MTCDs ($H = 25$).

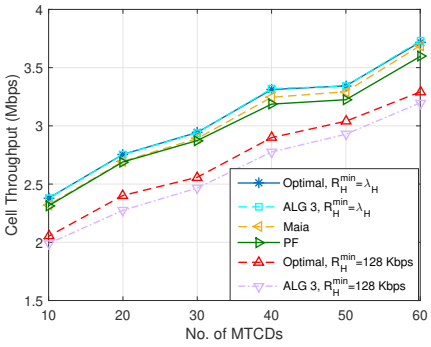
Fig. 2.4: Comparison of the proposed algorithms and other scheduling algorithms for Case 2.

validates the analytical analysis derived in Section 2.4.3.

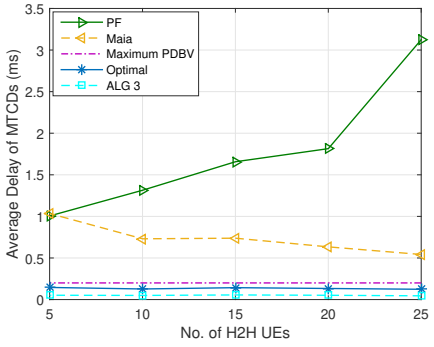
In Figure 2.3, we compare the performance of the optimal solution, using the BB algorithm, with that of Algorithm 2.1 and the other scheduling algorithms, for Case 1 scenario. The performance is measured in terms of aggregate cell throughput and average PDBV per MTCD in the cell. As apparent, the optimal solution and Algorithm 2.1 are identical in the performance at all the considered cases of number of users and values of R_u^{min}, $u \in \mathcal{H}$. This is due to the wideband CQI reporting considered in Case 1 which makes Algorithm 2.1 exhibit an optimal performance as



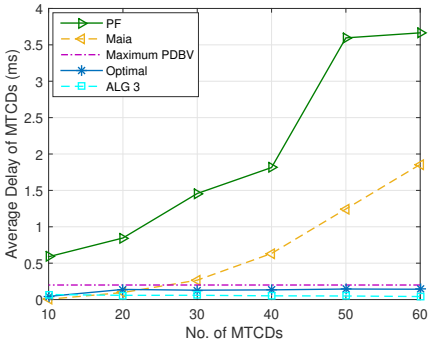
(a) Cell throughput vs. number of H2H UEs ($M = 40$).



(b) Cell throughput vs. number of MTCDs ($H = 15$).



(c) Average delay of MTCDs vs. number of H2H UEs ($M = 40$).



(d) Average delay of MTCDs vs. number of MTCDs ($H = 20$).

Fig. 2.5: Comparison of the proposed algorithms and other scheduling algorithms for Case 3.

discussed in Section 2.5.1.

In terms of cell aggregate throughput, the proposed algorithms, the optimal solution and Algorithm 2.1, exhibit the best performance compared to the PF and Maia algorithms. This is shown in Figs. 2.3(a) and 2.3(b). This is because the PF scheduler targets fair allocation on all the scheduled users including the MTCDs which are low data rate users. On the other side, the Maia scheduler subdivides the PRBs between the MTCDs and H2H UEs before scheduling. This impacts the H2H traffic, which contributes the major part of the cell throughput, by reducing the number of the avail-

able PRBs for it. However, the proposed algorithms allocate most of the resources to the H2H UEs to maximize their throughput considering the QoS requirements of the users. This concludes that the satisfaction of the QoS requirements of the users, as discussed below, is not at the expense of the cell throughput, but using a more efficient design.

The fulfillment of the QoS requirements of the MTCDs is apparent in Figs. 2.3(c) and 2.3(d) which show the average PDBV per MTCD in the cell. As the figures reveal, the proposed algorithms satisfy the cross-layer constraint at all cases. On the other hand, the PF scheduler does not consider any delay requirements for the users. The Maia algorithm satisfies the delay requirements at low numbers of devices, however, the violation of the delay requirements increases with the increase of the number of MTCDs.

Figure 2.4 compares the performance of the proposed algorithms with that of the other scheduling algorithms for Case 2. However Algorithm 2.2 is a suboptimal algorithm, it achieves an optimal throughput when $R_u^{min} = \lambda_u,\ u \in \mathcal{H}$. This is apparent in Figs. 2.4(a) and 2.4(b). This is due to the fact that the throughput of a certain queue is the same as the arrival rate if the service rate is greater than or equal to the arrival rate. Therefore, the maximum throughput can be achieved by guaranteeing a service rate that is greater than or equal to the arrival rate. Similar to Case 1, the proposed algorithms outperform the PF and Maia algorithms in terms of cell throughput and average PDBV per MTCD. The figure also reveals that the cell throughput is reduced when the minimum rate requirement of H2H UEs is adjusted to a value that is different than their arrival rates. This is because the H2H UEs contribute the larger part of the cell throughput. When they are served with a data rate that is less than their arrival rate, their throughput decreases, and consequently the aggregate cell throughput. In addition, the figure shows that the satisfaction of

the QoS requirements of the users, as in Figs. 2.4(c) and 2.4(d), is not at the expense of the cell throughput as Figs. 2.4(a) and 2.4(b) reveal, but using a more efficient design.

Figure 2.5 shows the performance of Algorithm 2.3 compared to that of the optimal solution and the other algorithms for Case 3. As it appears in the figure, Algorithm 2.3 achieves an optimal throughput when $R_u^{min} = \lambda_u,\ u \in \mathcal{H}$. However, the performance is suboptimal when $R_u^{min} = 128$ Kbps, $u \in \mathcal{H}$. The figure also reveals the superiority of the proposed algorithms in terms of cell throughput and average delay per MTCD.

2.7 Conclusions

In this chapter, we addressed the problem of resource allocation in an LTE cell that includes both critical-MTC and H2H communications. We formulated it as an optimization problem considering the QoS requirements of both types of communications in a cross-layer design perspective. We provided the methods to handle the optimization problems to get the optimal solution with a lower complexity. Moreover, more computationally-efficient algorithms have been proposed to implement the resource allocation process practically. The computational complexity of the proposed algorithms was investigated and discussed. The performance of the proposed methods and algorithms has also been evaluated analytically and by simulations. The simulations results validate the analytical analysis and reveal that the proposed methods outperform the other resource allocation algorithms from previous studies in the literature.

References

[1] H. Sundmaeker, P. Guillemin, P. Friess, and S. Woelfflé, "Vision and challenges for realising the internet of things," *Cluster of European Research Projects on the Internet of Things, European Commision*, 2010.

[2] M. Weyrich, J.-P. Schmidt, and C. Ebert, "Machine-to-machine communication," *IEEE Softw.*, vol. 31, no. 4, pp. 19–23, 2014.

[3] E. Dahlman, S. Parkvall, and J. Skold, *4G, LTE-advanced Pro and the Road to 5G*. Academic Press, 2016.

[4] ITU-R M.2083, *IMT vision - Framework and overall objectives of the future development of IMT for 2020 and beyond*, Sept. 2015.

[5] J. Sachs, G. Wikstrom, T. Dudda, R. Baldemair, and K. Kittichokechai, "5g radio network design for ultra-reliable low-latency communication," *IEEE Network*, vol. 32, no. 2, pp. 24–31, 2018.

[6] 3GPP TR 36.881 v14.0.0, *Study on latency reduction techniques for LTE (Release 14)*, June 2016.

[7] J. C. S. Arenas, T. Dudda, and L. Falconetti, "Ultra-low latency in next generation lte radio access," in *SCC 2017; 11th International ITG Conference on Systems, Communications and Coding; Proceedings of*. VDE, 2017, pp. 1–6.

[8] N. Afrin, J. Brown, and J. Y. Khan, "A delay sensitive lte uplink packet scheduler for m2m traffic," in *2013 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2013, pp. 941–946.

[9] A. E. Mostafa and Y. Gadallah, "A statistical priority-based scheduling metric for m2m communications in lte networks," *IEEE Access*, vol. 5, pp. 8106–8117, 2017.

[10] ——, "Uniqueness-based resource allocation for m2m communications in narrowband iot networks," in *Vehicular Technology Conference (VTC-Fall), 2017 IEEE 86th.* IEEE, 2017, pp. 1–5.

[11] A. Aijaz and A. H. Aghvami, "On radio resource allocation in lte networks with machine-to-machine communications," in *Vehicular Technology Conference (VTC Spring), 2013 IEEE 77th.* IEEE, 2013, pp. 1–5.

[12] A. Aijaz, M. Tshangini, M. R. Nakhai, X. Chu, and A.-H. Aghvami, "Energy-efficient uplink resource allocation in lte networks with m2m/h2h co-existence under statistical qos guarantees," *IEEE Transactions on Communications*, vol. 62, no. 7, pp. 2353–2365, 2014.

[13] M. Y. Abdelsadek, Y. Gadallah, and M. H. Ahmed, "An lte-based optimal resource allocation scheme for delay-sensitive m2m deployments coexistent with h2h users," in *2017 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, May 2017, pp. 139–144.

[14] M. K. Giluka, N. Rajoria, A. C. Kulkarni, V. Sathya, and B. R. Tamma, "Class based dynamic priority scheduling for uplink to support m2m communications in lte," in *Internet of Things (WF-IoT), 2014 IEEE World Forum on.* IEEE, 2014, pp. 313–317.

[15] A. G. Gotsis, A. S. Lioumpas, and A. Alexiou, "Analytical modelling and performance evaluation of realistic time-controlled m2m scheduling over lte cellular networks," *Transactions on Emerging Telecommunications Technologies*, vol. 24, no. 4, pp. 378–388, 2013.

[16] F. Ghavimi, Y.-W. Lu, and H.-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, 2017.

[17] C. She, C. Yang, and T. Q. Quek, "Cross-layer optimization for ultra-reliable and low-latency radio access networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 1, pp. 127–141, 2018.

[18] ——, "Radio resource management for ultra-reliable and low-latency communications," *IEEE Communications Magazine*, vol. 55, no. 6, pp. 72–78, 2017.

[19] C. Sun, C. She, C. Yang, T. Q. Quek, Y. Li, and B. Vucetic, "Optimizing resource allocation in the short blocklength regime for ultra-reliable and low-latency communications," *IEEE Transactions on Wireless Communications*, vol. 18, no. 1, pp. 402–415, 2019.

[20] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Transactions on Information Theory*, vol. 56, no. 5, p. 2307, 2010.

[21] W. Yang, G. Durisi, T. Koch, and Y. Polyanskiy, "Quasi-static multiple-antenna fading channels at finite blocklength," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 4232–4265, 2014.

[22] A. M. Maia, D. Vieira, M. F. de Castro, and Y. Ghamri-Doudane, "A fair qos-aware dynamic lte scheduler for machine-to-machine communication," *Computer Communications*, vol. 89, pp. 75–86, 2016.

[23] A. Azari and G. Miao, "Network lifetime maximization for cellular-based m2m networks," *IEEE Access*, vol. 5, pp. 18 927–18 940, 2017.

[24] A. Elhamy and Y. Gadallah, "Bat: A balanced alternating technique for m2m uplink scheduling over lte," in *2015 IEEE 81st Vehicular Technology Conference (VTC Spring)*. IEEE, 2015, pp. 1–6.

[25] 3GPP TR 38.913, *Study on scenarios and requirements for next generation access technologies, technical specification group radio access network*, Oct. 2016.

[26] P. Popovski, "Ultra-reliable communication in 5g wireless systems," in *5G for Ubiquitous Connectivity (5GU), 2014 1st International Conference on*. IEEE, 2014, pp. 146–151.

[27] N. Abu-Ali, A.-E. M. Taha, M. Salah, and H. Hassanein, "Uplink scheduling in lte and lte-advanced: Tutorial, survey and evaluation framework," *IEEE Communications surveys & tutorials*, vol. 16, no. 3, pp. 1239–1265, 2014.

[28] 3GPP R1-101211, *PUSCH Resource Allocation for Clustered DFT-Spread OFDM*, Feb. 2010.

[29] 3GPP TR 25.814, *Physical layer aspects for evolved Universal Terrestrial Radio Access (UTRA)*, v. 7.0.0, 2006.

[30] *Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures*, 3GPP TS 36.213, Std. v.9.2.0, (Release 9), June 2010.

[31] *Analysis on traffic model and characteristics for MTC and text proposal*, 3GPP R1-120056, Technical Report, TSG-RAN Meeting WG1#68, Dresden, Germany, Feb. 2012.

[32] A. G. Gotsis, A. S. Lioumpas, and A. Alexiou, "M2m scheduling over lte: Challenges and new perspectives," *IEEE Vehicular Technology Magazine*, vol. 7, no. 3, pp. 34–39, 2012.

[33] *Policy and charging control architecture*, 3GPP TS 23.203, Std. v.10.6.0, (Release 10), Mar. 2012.

[34] A. K. Erlang, "The theory of probabilities and telephone conversations," *Nyt Tidsskrift for Matematik B*, vol. 20, no. 33-39, p. 16, 1909.

[35] C. H. Papadimitriou and K. Steiglitz, *Combinatorial optimization: algorithms and complexity*. Courier Corporation, 1998.

[36] S. Haney, "Practical applications and properties of the exponentially modified gaussian (emg) distribution." Ph.D. dissertation, Drexel University, 2011.

[37] J. F. Shortle, M. J. Fischer, and P. H. Brill, "Waiting-time distribution of m/dn/1 queues through numerical laplace inversion," *INFORMS journal on Computing*, vol. 19, no. 1, pp. 112–120, 2007.

[38] C.-S. Chang and J. A. Thomas, "Effective bandwidth in high-speed digital networks," *IEEE Journal on Selected areas in Communications*, vol. 13, no. 6, pp. 1091–1100, 1995.

[39] Alberto Leon-Garcia, *Probability, Statistics, and Random Processes for Electrical Engineering*. Pearson Prentice Hall, 2008.

[40] *Further advancements for E-UTRA physical layer aspects*, 3GPP TR 36.814, 2010.

[41] W. Anchun, X. Liang, Z. Shidong, X. Xibin, and Y. Yan, "Dynamic resource management in the fourth generation wireless systems," in *Communication Technology Proceedings, 2003. ICCT 2003. International Conference on*, vol. 2. IEEE, 2003, pp. 1095–1098.

Chapter 3

A Two-Sided Matching Approach for Resource Allocation

3.1 Abstract

Cellular systems present one of the most suitable wireless technologies to efficiently serve critical Machine-Type Communications (MTC) that require strict Quality-of-Service (QoS) guarantees. Therefore, Ultra-Reliable and Low-Latency Communications (URLLC) is a target use case in the design of the upcoming generations of cellular networks. From the radio resource management perspective, guaranteeing such stringent QoS requirements in LTE networks is a challenging task, especially in the case of the coexistence of MTC with the Human-Type Communications (HTC). In this chapter, we address the resource allocation and scheduling problem of critical MTC that coexist with HTC in LTE. The optimization problem is formulated such that the overall system utility is maximized while fulfilling the different QoS demands of the two sets of users. Utilizing the effective bandwidth and effective capacity theories, a cross-layer design is developed to guarantee the QoS requirements of the critical

MTC. For a computationally-efficient solution of the problem, we formulate it as a two-sided matching process that can be used as a practical scheduling scheme. To this end, we analyze the convergence, stability, and computational complexity of the proposed methods. Results reveal the close-to-optimal performance of the matching-based scheduling scheme and its superiority to other existing techniques.

3.2 Introduction

One of the key features of smart devices, or intelligent machines, is their ability to be interconnected and remotely accessed. This gives rise to the Machine-to-Machine (M2M) type of communication in which there is no human intervention. M2M communications therefore enable the Internet of Things (IoT) and tactile Internet. From a communication technology point of view, there are two major categories of M2M. The first type is the massive Machine-Type Communications (MTC) which emphasizes the connectivity of a large number of low-cost and low-complexity smart objects. These devices are characterized by their low data rate and delay-tolerant traffic. The second category, which is the focus of this chapter, is the critical MTC which supports applications such as traffic safety, e-health, smart grid, emergency and disaster response, and industry automation. Data transmission of critical MTC requires high reliability, network availability and low end-to-end latency.

To guarantee such stringent Quality of Service (QoS) requirements, cellular networks present one of the most suitable technologies to serve critical MTC due to its ubiquitous coverage, advanced Radio Resource Management (RRM) procedures, and licensed spectrum availability. Ultra-Reliable and Low-Latency Communications (URLLC), along with the enhanced Mobile Broadband (eMBB) and massive MTC, are considered a major use case in the design of the upcoming generations of cellular

networks [1]. Besides the New Radio (NR), the evolution of the Long Term Evolution (LTE) is being standardized to achieve the fifth generation (5G) requirements in a backward-compatible manner [2]. Therefore, several enhancements and features have been proposed in Releases 14 and 15 of LTE to efficiently serve URLLC [3]. These features include fast uplink access on Medium Access Control (MAC), short transmission time intervals, and reduced processing time.

From the RRM perspective, guaranteeing the strict QoS requirements of the critical MTC is a challenging task [4]. This is due to the characteristics of MTC that are significantly different from those of the Human-to-Human (H2H) communications. Therefore, handling the critical MTC in the same way as traditional H2H real-time services is not efficient due to the different level of QoS requirements and packet size. The H2H traffic requires moderate latency and reliability guarantees compared to those of the critical MTC. In addition, H2H packet sizes are considerably larger than those of the critical MTC. Another major challenge is that the fulfilment of the stringent QoS demands of the critical MTC can impact the QoS of the H2H communications and the overall system utility. Therefore, one should consider all these factors in the design of RRM techniques.

The resource scheduling and allocation process is at the heart of the RRM procedures in LTE. It is crucial in optimizing the system performance as it affects the users' QoS performance and hence the overall system utility. In this chapter, we address the design of the scheduling process with the coexistence of both the critical MTC and Human-Type Communications (HTC) while considering the aforementioned aspects.

3.2.1 Related Work

To enable critical MTC in LTE, several studies consider the scheduling process of this type of communications. In [5] and [6], the authors propose prioritizing the MTC

devices (MTCDs) according to the statistical priority of the data to be transmitted. This is achieved by allowing the MTCDs to report a metric that indicates this information before data transmission. However, the current MAC protocol of the LTE standard needs to be changed to support those kinds of reports.

The authors in [7] cluster the MTCDs depending on their QoS requirements and transmission protocols. After that, they consider maximizing their aggregated data rate while satisfying minimum rate requirements. However, due to the fact that the data transmissions of critical MTCDs are identified by their low rate, maximizing their data rate is not efficient as will be elaborated on in Section 3.3. In addition, they do not consider the impact on the HTC users.

Further, several studies consider the partitioning of the resources between the MTC and HTC users before the scheduling process. In [8], the authors target a balance between the QoS requirements and the throughput of the MTCDs after splitting the resources according to the number of MTC and HTC users. The authors in [9] consider the buffer sizes of the users to split the resources. Then, the MTC traffic is scheduled separately based on fairness and their transmission deadlines. However, splitting the resources between the two types of traffic in such manner does not optimize the resources utilization at the system level.

Besides, several works investigate the resource allocation problem for the MTC coexistent with HTC without splitting the resources. The authors in [10], address the problem of maximizing the energy efficiency (expressed in bits per joule capacity of all users) while considering the delay requirements of the MTC traffic with statistical guarantees. However, the rate maximization for the MTC devices that transmit small-size packets negatively impacts that of the HTC users. In [11], we target maximizing the HTC sum-rate while fulfilling the latency constraints of the MTCDs. However, we examine a special case of the channels conditions and do not consider the com-

putational complexity of the proposed algorithms. Similarly, in [12], we maximize the aggregate HTC rate considering the different cases of channel condition feedback. However, in the general case, we consider average delay of the critical MTC as a QoS metric. In addition, we adopt $M/G/1$ queue model analysis to formulate the average delay constraint of the devices. In [13], we address the resource allocation problem targeting massive Multiple-Input Multiple-Output (MIMO) systems.

3.2.2 Paper Contributions

The principal contributions of the present chapter are as follows:

- The resource allocation for the critical MTC coexisting with HTC is formulated as a maximization problem to optimize the scheduling process. In the optimization problem, we maximize the HTC sum-rate and take into consideration the diverse QoS demands of the two different kinds of users. In this regard, a statistical QoS guarantee is used for the fulfillment of the demands of the critical MTC in the cell. More specifically, the probability of delay-bound violation (PDBV) of the critical MTCDs is restricted to be within a certain threshold that is determined for every device. This provides an accurate statistical guarantee for the latency and the packet losses resulting from queuing delay. As will be discussed in Section 3.3, this metric is considered more accurate than average delay that is adopted in [12] in the general scenario of channel feedback. In addition, we use the Effective Bandwidth (EB) [14] and Effective Capacity (EC) [15] theories to formulate the PDBV constraints in a cross-layer manner. This design considers both the buffer dynamics and the PHYsical (PHY) layer parameters of the devices. However, the formulated problem is combinatorial with high computational complexity. Therefore, we exploit the ergodicity of the service processes to simplify the formulated problem so that we can solve it

optimally with lower complexity.

- Moreover, we use the matching theory [16] to formulate the simplified resource allocation problem as a two-sided matching process. The proposed matching-based algorithms are less complex than those discussed in [13], since they target massive MIMO systems and are designed to consider the interference between users. This makes those algorithms are not computationally-efficient to be applied in the considered case of single antenna systems. The matching theory is a recently used technique in RRM in wireless networks that overcomes the limitations of the optimization and game theory [17]. It provides the mathematical framework for solving combinatorial problems in a computationally-efficient manner. Therefore, the designed algorithms can run in a polynomial time and can be used as a practical scheduling scheme. In this regard, the proposed matching-based scheme is analyzed from a practical perspective. Specifically, we discuss the convergence, stability, and computational complexity of the proposed methods. Besides, we implement extensive simulations to measure the performance of the matching-based scheduling scheme and prove that it approaches the optimal performance as will be discussed in Section 3.6.

The remainder of the chapter is organized as follows. In Section 3.3, we describe the system model and formulate the optimization problem after discussing the EB and EC concepts. Then, the problem is simplified and formulated as a two-sided matching process in Section 3.4. In Section 3.5, we analyze the proposed matching-based scheduling scheme from a practical point of view. Then, the results of the simulations implemented to evaluate the performance of the proposed scheme are discussed in Section 3.6. Finally, the chapter is concluded in Section 3.7.

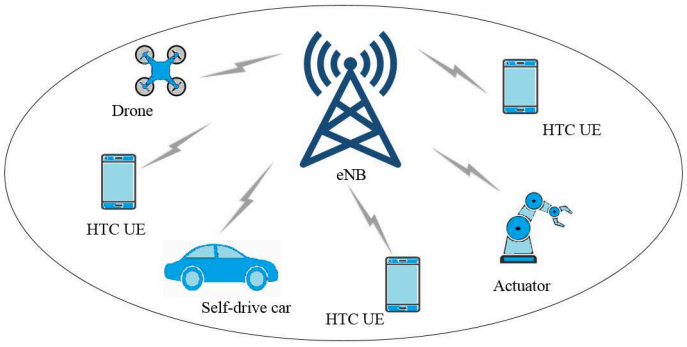


Fig. 3.1: An LTE cell with critical MTC coexistent with HTC.

Table 3.1: Summary of Notations of Chapter 3

Notation	**Definition**
$\mathcal{K}$	Set of PRBs
$\mathcal{K}_u$	PRBs allocated to uth user
$\mathcal{U}$	All users set
$\mathcal{H}$	Subset of HTC UEs
$\mathcal{M}$	Subset of MTCDs
K, U, H, M	Cardinalities of $\mathcal{K}, \mathcal{U}, \mathcal{H}, \mathcal{M}$
R_u	Data rate of user u
$R_{u,k}$	Data rate of user u on PRB k
$\bar{R}_u^{min}$	Minimum average rate of user u
q_u^{max}	Maximum no. of PRBs allocated to user u
$s_{u,k}$	Allocation indicator of PRB k to user u
D_u	Delay of user u
D_u^{max}	Delay-bound of user u
V_u^{max}	PDBV bound of user u
θ_u	QoS exponent
β_u	EB
α_u	EC

3.3 System Model and Problem Formulation

We consider the uplink scheduling process of critical MTC coexistent with HTC in a single LTE cell with a single eNB, as shown in Fig. 3.1. The system spectrum is divided into blocks of 180 KHz bandwidth each. Each of these blocks is called a physical resource block (PRB). Each user in the cell is allocated a subset of PRBs by the eNB scheduler to transmit its data for a duration of 1 ms, which is referred to as the transmission time interval (TTI). Therefore, the scheduling process is implemented every TTI to allocate the available PRBs to the users aiming at maximizing the utility of the system while fulfilling the QoS demands of the users.

Assume that the set of available PRBs is $\mathcal{K} = \{1, \cdots, k, \cdots, K\}$ and the set of active users to be served is $\mathcal{U} = \{1, \cdots, u, \cdots, U\}$. The set of users consists of two subsets, i.e., $\mathcal{U} = \mathcal{H} \cup \mathcal{M}$ where $\mathcal{H}$ and $\mathcal{M}$ are the subsets of HTC users and critical MTCDs, respectively. The cardinalities of the sets $\mathcal{K}, \mathcal{U}, \mathcal{H}, \mathcal{M}$ are K, U, H, M, respectively. We summarize the notations in Table 3.1.

The HTC User Equipment (UEs) are data-hungry devices, which means that serving them with higher data rates enhances their QoS. This consequently increases the system utility which depends on the aggregate throughput of the users in the cell. On the other hand, the MTCDs are characterized by their low data rate transmissions. However, they have critical delay requirements that need to be satisfied. This is because their packets can be useless if they arrive after their deadlines. Thus, given that the delay requirements are fulfilled, increasing the data rates of the MTCDs does not yield an improvement in their QoS. This means that simply considering the maximization of the data rates of the MTCDs will be at the cost of degrading those of the HTC UEs without actually improving the QoS of MTCDs.

From the previous discussion, we formulate the resource allocation optimization as a constrained maximization problem. The objective function, to be maximized in

this case, is the aggregate data rate of the HTC UEs. We formulate the constraints such that the QoS demands of both types of communications are fulfilled. Therefore, the optimization problem can be formulated as follows

$$\max_{\{\mathcal{K}_1,\cdots,\mathcal{K}_U\}\in\mathcal{K}} \sum_{u\in\mathcal{H}} R_u \tag{3.1}$$

$$\text{s.t. } \mathcal{K}_u \cap \mathcal{K}_{u'} = \varnothing,\ \forall u \neq u', u, u' \in \mathcal{U} \tag{3.1a}$$

$$\mathbb{E}\{R_u\} \geq \bar{R}_u^{min},\ \forall u \in \mathcal{H} \tag{3.1b}$$

$$\text{Cross-layer constraint},\ \forall u \in \mathcal{M} \tag{3.1c}$$

$$|\mathcal{K}_u| \leq q_u^{max},\ \forall u \in \mathcal{M}, \tag{3.1d}$$

where R_u is the data rate of the uth user over the subset of PRBs allocated to it, $\mathcal{K}_u \in \mathcal{K}$, and $\mathbb{E}\{\cdot\}$ is the expectation. Constraint (3.1a) is expressed to make sure that every PRB is assigned to only one user at most, while constraint (3.1b) is used to guarantee a minimum QoS for the HTC UEs by maintaining a minimum average data rate $\bar{R}_u^{min}$. However, the QoS of the MTCDs is considered by a cross-layer constraint in (3.1c) as will be discussed below. Finally, constraint (3.1d) is used to provide an upper-bound of the number of the assigned PRBs to a certain MTCD, where $|\mathcal{K}_u|$ is the cardinality of the PRBs subset allocated to the uth user. For example, in Release 13 of LTE, 6 PRBs is the maximum allowed number of PRBs that can be assigned to an MTCD.

To consider the delay constrains of critical MTC traffic, a cross-layer design is required to take into consideration both the buffer dynamics and the PHY layer parameters. A straightforward approach is to maintain the average delay below a certain level. However, some packets could be served with a low delay and others with a large delay which, in total, gives the required average delay. Therefore, many packets could be useless because they exceed their deadline while the average delay

constraint is still satisfied. Thus, considering the average delay can be inaccurate. Nevertheless, a more practical and accurate approach is to consider the probability of delay bound violation (PDBV) as a constraint, which is aligned with the QoS guarantees adopted in LTE [18]. This is because it provides a statistical guarantee for the latency and the packet losses resulting from queuing delay. That is, the cross-layer constraint in (3.1c) can be formulated as

$$\Pr[D_u \geq D_u^{max}] \leq V_u^{max},\ \forall u \in \mathcal{M}, \tag{3.2}$$

where D_u^{max} is the delay bound and V_u^{max} is the maximum tolerated PDBV for the uth MTCD. To express this constraint in terms of PHY layer parameters, we utilize the theories of large deviations, EB, and EC, as discussed in Sections 3.3.1 and 3.3.2.

In addition to using the cross-layer constraint in (3.2), we use an indicator matrix $\mathbf{S}$ which is a $U \times K$ binary matrix. Every element, $s_{u,k}$, of $\mathbf{S}$ indicates whether PRB k is allocated to user u. Therefore, the resulting equivalent optimization problem is

$$\max_{\mathbf{S}} \quad \sum_{u\in\mathcal{H}} \sum_{k=1}^{K} R_{u,k} s_{u,k} \tag{3.3}$$

$$\text{s.t.} \quad \sum_{u=1}^{U} s_{u,k} \leq 1,\ \forall k \in \mathcal{K} \tag{3.3a}$$

$$\mathbb{E}\{R_u\} \geq \bar{R}_u^{min},\ \forall u \in \mathcal{H} \tag{3.3b}$$

$$\Pr[D_u \geq D_u^{max}] \leq V_u^{max},\ \forall u \in \mathcal{M} \tag{3.3c}$$

$$\sum_{k=1}^{K} s_{u,k} \leq q_u^{max},\ \forall u \in \mathcal{M} \tag{3.3d}$$

$$s_{u,k} \in \{0,1\},\ \forall u \in \mathcal{U},\ k \in \mathcal{K}, \tag{3.3e}$$

where $R_{u,k}$ is the achievable data rate of the uth user over the kth PRB. Constraints (3.3a)–(3.3d) correspond to (3.1a)–(3.1d), respectively. In addition, we use the con-

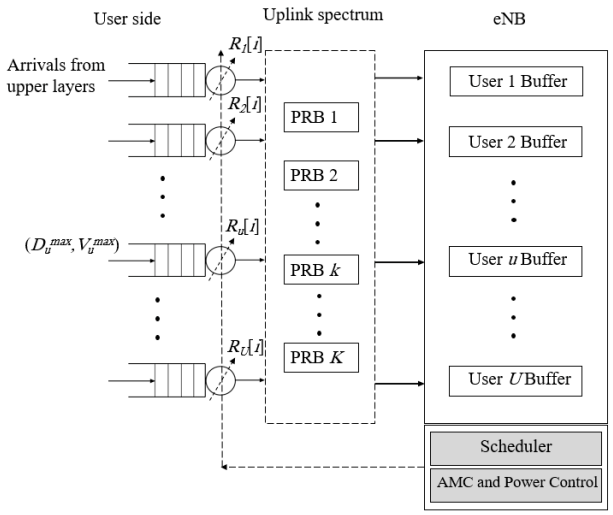


Fig. 3.2: Uplink cross-layer scheduling.

straint in (3.3e) to restrict the indicator variable $s_{u,k}$ to a binary value.

The achievable data rate of the uth user over the kth PRB, $R_{u,k}$, can be calculated from the signal-to-noise ratio (SNR), $\gamma_{u,k}$, which is given by

$$\gamma_{u,k} = \frac{P_{u,k}|h_{u,k}|^2}{N_0 B^{PRB}}, \tag{3.4}$$

where N_0 is the power spectral density (PSD) of the additive white Gaussian noise (AWGN), $B^{PRB} = 180$ KHz is the PRB bandwidth, $|h_{u,k}|^2$ is the channel power gain, and $P_{u,k}$ is the transmit power of user u over PRB k that is adjusted by the power control process. In LTE, a limited number of adaptive modulation and coding (AMC) schemes is used to adapt the link for every SNR level. Therefore, the achievable data

rates of the users on every PRB are discrete and can be calculated by

$$R_{u,k}(\gamma_{u,k}) = \begin{cases} r_0, & \gamma_{u,k} < \eta_0 \\ r_1, & \eta_0 \leq \gamma_{u,k} < \eta_1 \\ . & . \\ & \\ . & . \\ . & . \\ r_A, & \gamma_{u,k} \geq \eta_A \end{cases}, \tag{3.5}$$

where $\{\eta_j : \ j = 0, 1, \cdots, A\}$ are the thresholds of the SNR values for every AMC scheme, and r_j is the corresponding achievable rate. Fig. 3.2 shows the cross-layer scheduling process as implemented in the eNB to control the uplink transmission rate and the buffer dynamics, in addition to the used AMC and power control.

3.3.1 Effective Bandwidth and Effective Capacity

In this section, we define the EB and EC concepts. We then use these concepts to formulate the PDBV constraint in (3.2).

For an arrival process $\mathcal{A}(t)$, the EB can be defined as the minimum constant service rate that can serve $\mathcal{A}(t)$ with an ensured QoS exponent θ, which is given by [14]

$$\beta(\theta) = \lim_{t\to\infty} \frac{1}{t\theta} \ln \mathbb{E}\{e^{\theta \mathcal{A}(t)}\}. \tag{3.6}$$

Similarly, for a service process $\mathcal{S}(t)$, the maximum constant arrival rate that can be served by $\mathcal{S}(t)$ with an ensured QoS exponent θ, is known as the EC of the process

and can be calculated as [15]

$$\alpha(\theta) = -\lim_{t\to\infty} \frac{1}{t\theta} \ln \mathbb{E}\{e^{-\theta \mathcal{S}(t)}\}. \tag{3.7}$$

The QoS exponent θ models the decreasing rate of the queue length such that the PDBV is approximated and bounded as follows

$$\Pr[D(t) \geq D^{max}] \approx e^{-\theta \delta D^{max}} \leq V^{max}, \tag{3.8}$$

where δ is a parameter that depends on both of $\mathcal{A}(t)$ and $\mathcal{S}(t)$, and is defined as the rate at which the EC and EB curves intersect [19]. Therefore, a smaller θ indicates a less stringent QoS constraint while a larger θ implies a more stringent QoS requirement.

3.3.2 Formulation of the Cross-Layer PDBV Constraint

According to 3GPP [20], the Poisson process can be used to accurately characterize the traffic of MTCDs. Therefore, the approximation of the PDBV as in the EB theory is accurate as the results in [21] reveal. To use the EB and EC to guarantee the constraint in (3.2), we assume that the uth MTCD is serviced by a data rate of $R_u[l]$ in the lth TTI that has a period of T seconds. The data rate can vary at every TTI depending on the resource allocation policy. Therefore, the sequence $\{R_u[l]T, l = 1, 2, \cdots\}$ represents a discrete-time stationary and ergodic random process and $\mathcal{S}[t] = \sum_{l=1}^{t} R_u[l]T$ is its partial sum over the sequence $l = 1, 2, ..., t$ of TTIs. The sequence $\{R_u[l]T, l = 1, 2, \cdots\}$ is an uncorrelated process. Therefore, the EC defined in (3.7) for this service process, $\mathcal{S}[t]$, reduces to [22]

$$\alpha_u(\theta_u) = \frac{-1}{\theta_u} \ln \mathbb{E}\{e^{-\theta_u R_u[l]T}\}, \tag{3.9}$$

where θ_u is the guaranteed QoS exponent by the service process $\{R_u[l]T\}$ for the uth MTCD. That is,

$$\Pr[D_u \geq D_u^{max}] \approx e^{-\theta_u \delta_u D_u^{max}} \leq V_u^{max}. \tag{3.10}$$

Therefore, the guaranteed QoS exponent for the uth user can be calculated from the equality in (3.10) as

$$\theta_u = \frac{-\ln V_u^{max}}{\delta_u D_u^{max}}. \tag{3.11}$$

The satisfaction of the constraint in (3.10) requires that $\alpha_u(\theta_u) \geq \beta_u(\theta_u)$. That is, using (3.9), the cross-layer delay constraint becomes [1]

$$\frac{-1}{\theta_u} \ln \mathbb{E}\{e^{-\theta_u R_u[l]T}\} \geq \beta_u. \tag{3.12}$$

Given (3.12) as the constraint in place of (3.3c), the problem in (3.3) can be modeled as a Binary Nonlinear Program (BNLP) which can be solved using algorithms such as the Branch and Bound (BB). Nevertheless, the BNLP is a combinatorial problem that is NP-hard [23]. Therefore, the computational complexity of the optimal solution is exponential, as will be discussed in Section 3.5.3, and cannot be used in the scheduling process that is required to be executed in intervals of 1 ms each in LTE. This motivates the use of the matching-based solutions as discussed in Section 3.4.

[1] $\beta_u(\theta_u)$ is written as β_u to simplify the equations.

3.4 Matching-Based Scheduling

In this section, we first simplify the problem to a Binary Linear Program (BLP) to calculate the optimal solution with a reduced complexity. Also, this enables us to formulate the problem as a two-sided matching process as in Section 3.4.2. After that, we propose a matching-based scheduling scheme in Section 3.4.3.

3.4.1 BLP Problem Simplification

The nonlinearity of the optimization problem in (3.3) results from the constraints in (3.3b) and (3.3c). Therefore, we derive equivalent linear constraints by calculating the instantaneous data rates that guarantee the satisfaction of those conditions as in Theorem 3.4.1.

Theorem 3.4.1. *The EC constraint in (3.12) is equivalent to the following instantaneous data rate constraint at the lth TTI:*

$$R_u[l] \geq \frac{1}{\theta_u T} \ln\left(l e^{-\theta_u \beta_u} - (l-1)\psi_u^{avg}[l-1]\right), \ \forall u \in \mathcal{M} \tag{3.13}$$

where,

$$\psi_u^{avg}[l] = \begin{cases} \frac{\psi_u[l]+(l-1)\psi_u^{avg}[l-1]}{l}, & l \geq 2 \\ \psi_u[l], & l = 1 \end{cases}, \tag{3.14}$$

$$\psi_u[l] = e^{-\theta_u R_u[l] T}. \tag{3.15}$$

In addition, the average rate constraint in (3.3b) is equivalent to

$$R_u[l] \geq l\bar{R}_u^{min} - (l-1)R_u^{avg}[l-1], \ \forall u \in \mathcal{H} \tag{3.16}$$

where,

$$R_u^{avg}[l] = \begin{cases} \frac{R_u[l]+(l-1)R_u^{avg}[l-1]}{l}, & l \geq 2 \\ R_u[l], & l = 1 \end{cases}. \tag{3.17}$$

Proof. To prove (3.13), we define $\psi_u[l]$ as in (3.15) and $\psi_u^{avg}[l]$ as the cumulative moving average (CMA) of $\psi_u[l]$ at the lth TTI which can be derived as in (3.14). This acts as the estimation of $\mathbb{E}\{\psi_u[l]\}$ because of the ergodicity of the random process $\{R_u[l]T\}$. Therefore, constraint (3.12) can be expressed as

$$\psi_u^{avg}[l] \leq e^{-\theta_u \beta_u}. \tag{3.18}$$

Using (3.14), (3.18) can be rearranged as

$$\psi_u[l] = e^{-\theta_u^{min} R_u[l]T} \leq l e^{-\theta_u \beta_u} - (l-1)\psi_u^{avg}[l-1]. \tag{3.19}$$

Solving (3.19) for $R_u[l]$, which is the instantaneous data rate, gives (3.13).

Similarly, define $R_u^{avg}[l]$ as the CMA of $R_u[l]$ at the lth TTI which is calculated by (3.17). This is the estimation of $\mathbb{E}\{R_u[l]\}$ due to its ergodicity. Rearranging gives (3.16) directly. □

Accordingly, using the equivalent constraints in (3.13) and (3.16) in place of (3.3c) and (3.3b), respectively, converts the BNLP problem in (3.3) into a BLP. Therefore,

the equivalent resource allocation problem at the lth TTI can be formulated as follows

$$\max_{\mathbf{S}} \quad \sum_{u\in\mathcal{H}}\sum_{k=1}^{K} R_{u,k}[l]s_{u,k} \tag{3.20}$$

$$\text{s.t.} \quad R_u[l] \geq R_u^{min}[l],\ \forall u \in \mathcal{U} \tag{3.20a}$$

$$\text{Constraint (3.3a)}$$

$$\text{Constraint (3.3d)}$$

$$\text{Constraint (3.3e)},$$

where $R_u^{min}[l]$ is the minimum instantaneous data rate required by the uth user at the lth TTI to satisfy its QoS requirements. This is the left hand side of (3.13) for $u \in \mathcal{M}$ and of (3.16) for $u \in \mathcal{H}$.

The BLP problem in (3.20) can be solved to get the optimal solution with reduced complexity compared to the BNLP problem in (3.3). Moreover, it enables us to formulate the scheduling problem as a two-sided matching process which yields a stable matching in a polynomial time as discussed below.

3.4.2 Two-Sided Matching Formulation

The optimization problem in (3.20) can be formulated as a two-sided matching scheme by considering the sets of users and PRBs, $\mathcal{U}$ and $\mathcal{K}$, as two sets of agents. Every PRB k from the set $\mathcal{K}$ seeks to match to an agent u from the set $\mathcal{U}$ such that its utility is maximized. Accordingly, it has a list of preference, $\mathcal{P}(k)$, in which it orders all the users $u \in \mathcal{U}$ based on its preference. This means that a user u' is preferred than user u'', written as $u' \succ_k u''$, if u' precedes u'' in the preference list of k, $\mathcal{P}(k)$. In a similar manner, every user $u \in \mathcal{U}$ has a preference over the PRBs $k \in \mathcal{K}$ listed in $\mathcal{P}(k)$. The preferences of all users and PRBs are transitive, i.e., if $k' \succ_u k''$ and $k'' \succ_u k'''$, then

$k' \succ_u k'''$. As will be discussed in Section 3.4.3, the utility functions of the agents are constructed in a way such that the matching of the agents maximizes the HTC sum-rate and satisfies the QoS demands of all users.

To represent the objective function and the constraints of the optimization problem in (3.20), we design the mapping function of the matching process with the following properties.

Definition 3.4.2. *The mapping function of the matching process, $\mu(\cdot)$, maps the agents from the set $\mathcal{U} \cup \mathcal{K}$ into the set $\mathcal{U} \cup \mathcal{K}$ with the following properties:*

(i) $\mu(u) \subseteq \mathcal{K},\ \forall u \in \mathcal{U}$

(ii) $\mu(k) \in \mathcal{U},\ \forall k \in \mathcal{K}$

(iii) $|\mu(u)| \geq q_u^{min},\ \forall u \in \mathcal{H}$

(iv) $q_u^{min} \leq |\mu(u)| \geq q_u^{max},\ \forall u \in \mathcal{M}$

(v) $k \in \mu(u)$ *if and only if* $\mu(k) = u$

Properties (i) and (ii) represent the constraint of the cardinality of the match set of every user and PRB, respectively. The match set of every user can contain more than one PRB, however, that of PRBs can only have one PRB. Properties (iii) and (iv) are equivalent to the constraints of minimum rate and maximum number of PRBs for the users by restricting the quota of every one of them. The minimum quota, q_u^{min}, is the cardinality of the set $\mu(u)$ that fulfill the minimum rate constraint of user u. Finally, property (v) ensures that a certain PRB can be in the match list of a user if and only if that user is the match of that PRB.

Algorithm 3.1 Proposed Scheduling Scheme Using Matching

Step 1: Initial setup

Calculate the required rate for all $u \in \mathcal{U}$ at the lth TTI using (3.13) and (3.16).

Construct the preference lists of all $k \in \mathcal{K}$ over $u \in \mathcal{U}$ based on $R_{u,k}$ such that the $\mathcal{H}$ users are more preferred than the $\mathcal{M}$ users.

Step 2: Phase 1 matching μ^{P1}

$\mathcal{U}^{P1} \leftarrow \mathcal{H}$, $\mathcal{K}^{P1} \leftarrow \mathcal{K}$

Construct the preference lists of all $u \in \mathcal{U}^{P1}$ over all $k \in \mathcal{K}^{P1}$ according to $R_{u,k}$.

Match the agents in $\mathcal{U}^{P1}$ and $\mathcal{K}^{P1}$ with relaxed minimum quotas ($q_u^{min} = 0,\ \forall u \in \mathcal{U}^{P1}$) using Gale-Shaply algorithm with k proposing.

Step 3: Setup of Phase 2

$\mathcal{U}^{sat} \leftarrow \{u \in \mathcal{U}^{P1} : R_u(\mu^{P1}(u)) \geq R_u^{min}[l]\}$

$\mathcal{U}^{unsat} \leftarrow \{u \in \mathcal{U}^{P1} : R_u(\mu^{P1}(u)) < R_u^{min}[l]\}$

for all $u \in \mathcal{U}^{sat}$ **do**

  Sort u's matched set of PRBs $\mu^{P1}(u)$ according to its preference list $\mathcal{P}(u)$ and match it only to the subset of PRBs that satisfy $R_u^{min}[l]$ (according to $\mathcal{P}(u)$) and put the remaining PRBs into the set $\mathcal{K}^{P2}$.

end for

Step 4: Phase 2 Matching μ^{P2}

$\mathcal{U}^{P2} \leftarrow \mathcal{M} \cup \mathcal{U}^{unsat}$

Construct the preference list of every $u \in \mathcal{U}^{P2}$ on every $k \in \mathcal{K}^{P2}$ based on the rate on every PRB of them by the old match in μ^{P1}, i.e., $k \succeq_u k'$ if $R_{\mu^{P1}(k),k} \geq R_{\mu^{P1}(k'),k'}$.

Modify the preference list of every $k \in \mathcal{K}^{P2}$ on every $u \in \mathcal{U}^{P2}$ such that $\mathcal{H}$ users are no longer preferred than $\mathcal{M}$ users.

Set the status of every $k \in \mathcal{K}^{P2}$ to 1, where only PRBs with status 1 still want to propose.

while any PRB's status is 1 **do**

  Determine the proposal of the first k. Suppose that its preferred user is u^*.

  if $R_{u^*,\mu^{P2}(u^*)} < R_{u^*}^{min}$ **and** $|\mu^{P2}(u^*)| \leq q_u^{max}$ **then**

    Match k to u^*.

  else

    Let u^* select the preferred subset of PRBs $\mathcal{C}$ from $\{k\} \cup \mu^{P2}(u^*)$ such that $R_{u^*,\mathcal{C}} \geq R_{u^*}^{min}$.

    Match u^* to $\mathcal{C}$ and reject the other PRBs.

    Set the status of every accepted PRB in $\mathcal{C}$ to 0.

    Remove u^* from the preference lists of the rejected PRBs and update their status to 1 if their preference lists are not empty.

  end if

end while

Rematch the PRBs that are not matched in μ^{P2} to their old match in μ^{P1}.

Step 5: Test of feasibility

if QoS of any $u \in \mathcal{U}^{P2}$ is not fulfilled **then**

  Matching is not feasible. **Stop**

end if

3.4.3 Matching-Based Scheduling Algorithm

To match the agents as described in Section 3.4.2, we propose Algorithm 3.1. Therefore, this algorithm can be used as a practical scheduling scheme that sub-optimally solves the problem in (3.20), and therefore that in (3.3), with reduced complexity relative to that of the optimal solution. This matching process can be modeled as a one-to-many problem that is similar to the Hospitals-Residents with lower and higher quota problem [24]. However, in this problem, the lower quota of user $u \in \mathcal{U}$ is determined using its minimum data rate constraint which depends on the matched set of PRBs, $\mu(u)$. Also, the HTC users have no higher quota bounds.

Algorithm 3.1 can be summarized as follows. We match the agents in $\mathcal{U}$ and $\mathcal{K}$ in two steps, referred to as Phase 1 and Phase 2. The preference list of every PRB $k \in \mathcal{K}$ is constructed according to the data rate of every user on it. That is,

$$u \succeq_k u' \Longleftrightarrow R_{u,k} \geq R_{u',k}, \tag{3.21}$$

such that the HTC users are more preferred than the MTC ones. Similarly, the preference list of every HTC user is based on the achievable data rates by this user on every PRB. However, the preference list of every MTC user is determined after Phase 1 of matching.

In Phase 1, we start by matching the HTC users, $u \in \mathcal{H}$, with the PRBs set $\mathcal{K}$ without lower quota bounds. For this purpose, we use the many-to-one Gale-Shapley algorithm [25], with the PRBs being the proposing agents. After the matching, we determine the HTC users that are satisfied in Phase 1 and the others that are not. Also, the PRBs that are allocated to the satisfied HTC users and more than the required ones, are determined as in Step 3 in Algorithm 3.1. In Phase 2, the MTC users and the HTC ones that have unsatisfied minimum rates, are considered for

matching. For these users, the remaining subset of PRBs is considered such that the minimum HTC data rate loss is achieved. Therefore, the preference lists of the users in Phase 2 are constructed based on the rates achieved by the satisfied HTC users in Phase 1 on the subset of PRBs considered in Phase 2. Consequently, at the end of Phase 2, the sum-rate of the HTC users is maximized and the QoS demands of the users are satisfied.

3.5 Analysis of the Matching-Based Scheduling

In this section, we analyze the convergence, stability, and complexity of the proposed matching-based scheduling algorithm.

3.5.1 Convergence

The convergence of the matching in Algorithm 3.1 can be analyzed as follows.

Lemma 3.5.1. *The proposed matching in Algorithm 3.1 converges to a matching μ^* after a limited number of iterations.*

Proof. As indicated in Algorithm 3.1, the PRBs are the proposing agents in both phases of the matching algorithm. Every PRB $k \in \mathcal{K}$ proposes to the first user $u \in \mathcal{U}$ in its preference list $\mathcal{P}(k)$. If it is not accepted by that user, it removes it from $\mathcal{P}(k)$ and proposes to the next user in the list until it is accepted by a user or the list $\mathcal{P}(k)$ is empty. The list $\mathcal{P}(k)$ of every PRB is finite due to the fact that the number of users is finite. As a result, the number of proposals and iterations is finite. Consequently, Algorithm 3.1 converges to a matching μ^* after a limited number of iterations. □

3.5.2 Stability

The considered matching problem falls in the two-sided matching with two-sided preferences category. In this category, the optimality criterion is the stability of the matching. To define and analyze the stability of the considered matching algorithm, we first define the blocking pair concept as follows.

Definition 3.5.2. *A pair (u', k') is called a blocking pair for the matching μ if:*

1. $\mu(k') \neq u',\ k' \notin \mu(u')$,
2. $u' \succ_{k'} \mu(k')$,
3. $\mathcal{C} \succ_{u'} \mu(u')$, *where* $\mathcal{C} \subseteq \{k'\} \cup \mu(u'),\ k' \in \mathcal{C}$, *and*
4. *the quota bounds of u' and $\mu(k')$ will still be satisfied if k' is matched to u'.*

Given the definition of a blocking pair, we can define the stability of the matching algorithm as follows [16].

Definition 3.5.3. *The matching process as in Definition 3.4.2 is stable if it admits no blocking pair.*

Based on Definitions 3.5.2 and 3.5.3, we can analyze the stability of the proposed matching scheme as in the following theorem.

Theorem 3.5.4. *Based on the definition of the stability as in Definition 3.5.3, the proposed matching algorithm converges to a stable matching μ^*.*

Proof. To prove the stability of the matching algorithm, we assume that Algorithm 3.1 has converged to a matching μ^* and there is a pair (u', k') where $\mu^*(k') \neq u',\ k' \notin \mu^*(u')$, $u' \succ_{k'} \mu^*(k')$, and the quota bounds of u' and $\mu^*(k')$ will still be satisfied if k' is matched to u'. This means that k' has proposed to u' before its final match

$\mu^*(k')$ at a certain iteration j in the matching and has been rejected. However, $\mu^{(j+1)}(u') \succeq_{u'} \mu^{(j)}(u')$, i.e., the matching of user u' at the $j+1$ iteration is preferred or at least as preferred as that at the jth iteration. This means that the subset of PRBs matched to user u' at the final iteration of matching is the most preferred subset for it. Thus, $\mathcal{C} \not\succ_{u'} \mu^*(u')$, where $\mathcal{C} \subseteq \{k'\} \cup \mu^*(u'),\ k' \in \mathcal{C}$. Therefore, according to Definition 3.5.2, this pair is not a blocking pair. Consequently, no blocking pair exists for the matching μ^*. Therefore, the matching μ^* is a stable matching. □

3.5.3 Computational Complexity

Due to the combinatorial structure of the scheduling optimization problem, the computational complexity of the optimal solution is exponential. This is considering that the worst case of the computational complexity of the BB technique can reach that of the exhaustive search. In this section, we investigate the computational complexity of the matching-based scheduling scheme and prove that it can run in a polynomial time. To this end, we derive the worst case computational complexity of the steps of Algorithm 3.1 in terms of the big-O notation. Then, the total computational complexity is approximated by the largest component.

The computational complexity of the steps of Algorithm 3.1 can be derived as:

- $\mathcal{O}(KU^2)$ for Step 1,
- $\mathcal{O}(HK^2)$ for Step 2,
- $\mathcal{O}(K^2)$ for Step 3, and
- $\mathcal{O}((U-1)(K-1)^2)$ for Steps 4–5.

Consequently, the computational complexity of the proposed scheduling scheme is $\mathcal{O}((U-1)(K-1)^2) \approx \mathcal{O}(UK^2)$. This shows how the computational complexity is

Table 3.2: Simulation Parameters of Chapter 3

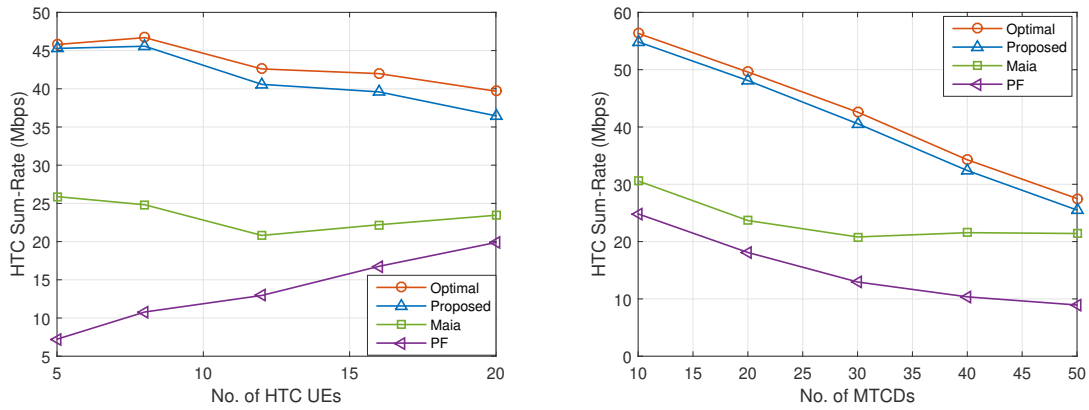
Parameter	**Value**
Cell radius	500 m
eNBs	1
Simulation period	1000 TTIs
Runs	20
Path loss	$128.1 + 37.6 \log_{10}(d)$, d in km [26]
Shadowing std	8 dB
Transmitter power	15 dBm
Noise PSD	-174 dBm/Hz
Noise figure	18 dB
Bandwidth	20 MHz
PRBs (K)	100
HTC arrival rate ($\lambda_u,\ u \in \mathcal{H}$)	$64, 128, 192, 256$ kbps
MTC arrival rate ($\lambda_u,\ u \in \mathcal{M}$)	$10, 20, 30, 40$ kbps [9]
Distribution of λ_u	Uniform
Delay-bound ($D_u^{max},\ u \in \mathcal{M}$)	0.3 ms
PDBV threshold ($V_u^{max},\ u \in \mathcal{M}$)	10^{-1}
$\bar{R}_u^{min},\ u \in \mathcal{H}$	λ_u

reduced using Algorithm 3.1 to a large extent relative to that of the optimal solution. Moreover, in Section 3.6, we show how close its performance is to that of the optimal solution. This enables Algorithm 3.1 to be used as a practical resource allocation scheme.

3.6 Results

In this section, we evaluate the performance of the proposed scheduling technique using simulations. The cell aggregate HTC achievable data rate is considered as the metric to measure the system utility. However, to check the satisfaction of the QoS demands of the MTCDs in the cell, we consider the average PDBV per MTCD as the performance metric.

In terms of these metrics, we compare the performance of the proposed scheduling



(a) HTC sum-rate versus number of HTC UEs ($M = 30$).

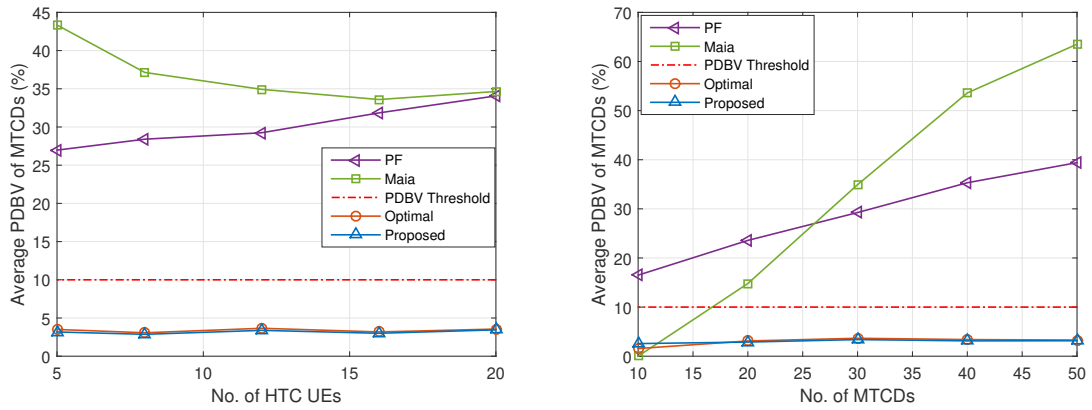
(b) HTC sum-rate versus number of MTCDs ($H = 12$).

Fig. 3.3: Aggregate HTC achievable data rate.

algorithm with that of the optimal solution of the problem in (3.20), the Proportional Fairness (PF) scheduling technique [27], and the first scheduling algorithm in [9], referred to as Maia algorithm. In [9], after splitting the resources between the MTC and HTC users, the HTC users are scheduled by the PF algorithm. However, the scheduling of the MTCDs considers their maximum delay and fairness. For the values of the parameters of this scheduling algorithm, we use the same as that in [9].

We consider a single LTE cell with a single eNB that serves a set of HTC users and critical MTCDs. The users are uniformly distributed within a circle of 500 m radius. The generated traffic is Poisson with arrival rates that are picked up randomly from the values shown in Table 3.2 along with other simulation parameters.

Figure 3.3 shows the HTC aggregate rate versus the number of users for the four scheduling algorithms. As the figure reveals, the proposed algorithm approaches the optimal solution in this metric, which is the objective function that is maximized in the optimization problem. However, the gap between them widens with the increase of the number of HTC users. Fig. 3.3(a) also shows that the HTC aggregate data rate



(a) Average PDBV versus number of HTC UEs ($M = 30$).

(b) Average PDBV versus number of MTCDs ($H = 12$).

Fig. 3.4: Average PDBV of the MTCDs.

increases by increasing the number of the HTC users. However, it starts to degrade after a certain limit. This is due to the minimum rate requirements of the HTC users that need to be satisfied before maximizing their aggregate data rate. Similarly, the fulfillment of the QoS demands of the MTCDs impacts the data rate of the HTC UEs as shown in Fig. 3.3(b). On the other hand, the PF scheduler assigns the resources in a fair way to all users without maximizing the rate of a certain subset. Moreover, it does not consider the fulfillment of the QoS demands of the users. Therefore, it allocates excess PRBs to the MTCDs which decreases the sum rate of the HTC UEs. For the Maia scheduler, the splitting of the resources between the HTC and MTC users before allocation is not optimal. This directly impacts the data rate of the HTC UEs due to the fact that excess PRBs given to the MTCDs are at the expense of the number of available PRBs for the HTC UEs.

The average PDBV of the critical MTCDs is plotted in Fig. 3.4 versus the number of users for the considered scheduling techniques. Again, the proposed algorithm is very close to the optimal solution. Both algorithms fulfill the QoS demands of the

MTCDs for any case of number of users. However, the PF scheduler does not consider any latency requirements and only respects the average data rate of the users. On the other hand, the Maia scheduler addresses the absolute deadlines of the MTCDs and does not consider their PDBV. Therefore, this only works in the cases of low numbers of MTCDs. However, when the number of MTCDs increases, it violates the statistical bound, as shown in Fig. 3.4(b).

3.7 Conclusions

In this chapter, we formulate the resource allocation and scheduling problem for a mix of HTC and MTC traffic sources such that the aggregate rate of the HTC is maximized while satisfying their QoS demands. The goal is to maximize the overall system throughput while fulfilling the QoS requirements of the critical MTC using a cross-layer design that is based on the effective bandwidth and effective capacity theories. This algorithm is intended for practical scheduling purposes. Therefore, we formulate the problem as a two-sided matching process which reduces the computational complexity significantly. The simulation experiments show that the performance of the proposed algorithm is very close to that of the optimal solution. Moreover, we analyze the proposed scheme from the practical perspective. Furthermore, simulations show that the proposed matching-based scheduler clearly outperforms the classical algorithms as well as the most significant techniques from the previous studies.

References

[1] ITU-R M.2083, *IMT vision - Framework and overall objectives of the future development of IMT for 2020 and beyond*, Sept. 2015.

[2] C. Hoymann, D. Astely, M. Stattin, G. Wikstrom, J.-F. Cheng, A. Hoglund, M. Frenne, R. Blasco, J. Huschke, and F. Gunnarsson, "Lte release 14 outlook," *IEEE Communications Magazine*, vol. 54, no. 6, pp. 44–49, 2016.

[3] J. C. S. Arenas, T. Dudda, and L. Falconetti, "Ultra-low latency in next generation lte radio access," in *SCC 2017; 11th International ITG Conference on Systems, Communications and Coding; Proceedings of.* VDE, 2017, pp. 1–6.

[4] C. She, C. Yang, and T. Q. Quek, "Radio resource management for ultra-reliable and low-latency communications," *IEEE Communications Magazine*, vol. 55, no. 6, pp. 72–78, 2017.

[5] A. E. Mostafa and Y. Gadallah, "A statistical priority-based scheduling metric for m2m communications in lte networks," *IEEE Access*, vol. 5, pp. 8106–8117, 2017.

[6] ——, "Uniqueness-based resource allocation for m2m communications in narrowband iot networks," in *Vehicular Technology Conference (VTC-Fall), 2017 IEEE 86th.* IEEE, 2017, pp. 1–5.

[7] F. Ghavimi, Y.-W. Lu, and H.-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, 2017.

[8] A. Elhamy and Y. Gadallah, "Bat: A balanced alternating technique for m2m uplink scheduling over lte," in *2015 IEEE 81st Vehicular Technology Conference (VTC Spring).* IEEE, 2015, pp. 1–6.

[9] A. M. Maia, D. Vieira, M. F. de Castro, and Y. Ghamri-Doudane, "A fair qos-aware dynamic lte scheduler for machine-to-machine communication," *Computer Communications*, vol. 89, pp. 75–86, 2016.

[10] A. Aijaz, M. Tshangini, M. R. Nakhai, X. Chu, and A.-H. Aghvami, "Energy-efficient uplink resource allocation in lte networks with m2m/h2h co-existence under statistical qos guarantees," *IEEE Transactions on Communications*, vol. 62, no. 7, pp. 2353–2365, 2014.

[11] M. Y. Abdelsadek, Y. Gadallah, and M. H. Ahmed, "An lte-based optimal resource allocation scheme for delay-sensitive m2m deployments coexistent with h2h users," in *2017 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, May 2017, pp. 139–144.

[12] ——, "Optimal cross-layer resource allocation for critical mtc traffic in mixed lte networks," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 6, pp. 5944–5956, June 2019.

[13] ——, "Matching-based resource allocation for critical mtc in massive mimo lte networks," *IEEE Access*, vol. 7, pp. 127 141–127 153, September 2019.

[14] C.-S. Chang and J. A. Thomas, "Effective bandwidth in high-speed digital networks," *IEEE Journal on Selected areas in Communications*, vol. 13, no. 6, pp. 1091–1100, 1995.

[15] D. Wu and R. Negi, "Effective capacity: a wireless link model for support of quality of service," *IEEE Transactions on wireless communications*, vol. 2, no. 4, pp. 630–643, 2003.

[16] M. David, *Algorithmics Of Matching Under Preferences*. World Scientific, 2013, vol. 2.

[17] Y. Gu, W. Saad, M. Bennis, M. Debbah, and Z. Han, "Matching theory for future wireless networks: fundamentals and applications," *IEEE Communications Magazine*, vol. 53, no. 5, pp. 52–59, May 2015.

[18] *Policy and charging control architecture*, 3GPP TS 23.203, Std. v.10.6.0, (Release 10), Mar. 2012.

[19] J. Tang and X. Zhang, "Cross-layer-model based adaptive resource allocation for statistical qos guarantees in mobile wireless networks," *IEEE Transactions on Wireless Communications*, vol. 7, no. 6, 2008.

[20] *Analysis on traffic model and characteristics for MTC and text proposal*, 3GPP R1-120056, Technical Report, TSG-RAN Meeting WG1#68, Dresden, Germany, Feb. 2012.

[21] G. L. Choudhury, D. M. Lucantoni, and W. Whitt, "Squeezing the most out of atm," *IEEE transactions on communications*, vol. 44, no. 2, pp. 203–217, 1996.

[22] J. Tang and X. Zhang, "Quality-of-service driven power and rate adaptation over wireless links," *IEEE Transactions on Wireless Communications*, vol. 6, no. 8, 2007.

[23] C. H. Papadimitriou and K. Steiglitz, *Combinatorial optimization: algorithms and complexity*. Courier Corporation, 1998.

[24] K. Hamada, K. Iwama, and S. Miyazaki, "The hopitals/residents problem with quota lower bounds," in *MATCH-UP 2008: Matching Under Preferences, satellite workshop of ICALP 2008*. IEEE, 2008, pp. 55–66.

[25] D. Gale and L. S. Shapley, "College admissions and the stability of marriage," *The American Mathematical Monthly*, vol. 69, no. 1, pp. 9–15, 1962.

[26] *Further advancements for E-UTRA physical layer aspects*, 3GPP TR 36.814, 2010.

[27] W. Anchun, X. Liang, Z. Shidong, X. Xibin, and Y. Yan, "Dynamic resource management in the fourth generation wireless systems," in *Communication Technology Proceedings, 2003. ICCT 2003. International Conference on*, vol. 2. IEEE, 2003, pp. 1095–1098.

Chapter 4

Resource Allocation in Massive MIMO LTE

4.1 Abstract

Supporting critical Machine-Type Communications (MTC) in addition to Human-Type Communications (HTC) is a major target for LTE networks to fulfill the 5G requirements. However, guaranteeing a stringent Quality-of-Service (QoS) for MTC, in terms of latency and reliability, while not sacrificing that of HTC is a challenging task from the radio resource management perspective. In this chapter, we optimize the resource allocation process through exploiting the additional degrees of freedom introduced by massive Multiple-Input Multiple-Output (MIMO) techniques. We utilize the effective bandwidth and effective capacity concepts to provide statistical guarantees for the QoS, in terms of probability of delay-bound violation, of critical MTC in a cross-layer design manner. In addition, we employ the matching theory to solve the formulated combinatorial problem with much lower computational complexity compared to that of the global optimal solution so that the proposed scheme can be used

in practice. In this regard, we analyze the computational complexity of the proposed algorithms and prove their convergence, stability and optimality. The results of extensive simulations that we performed show the ability of the proposed matching-based scheme to satisfy the strict QoS requirements of critical MTC with no impact on those of HTC. In addition, the results show a close-to-global optimal performance while outperforming other algorithms that belong to different scheduling strategies in terms of the adopted performance indicators.

4.2 Introduction

In order to accommodate all communicating elements to be connected to the network and form the Internet of Things (IoT), the evolution of the communication networks to incorporate Machine-Type Communications (MTC) in addition to Human-Type Communications (HTC) has become inevitable. MTC can be categorized into two major classes, massive MTC and critical MTC. The former is about connecting a massive number of low-complexity and low-cost devices such as sensors and wearables. It supports the IoT applications that require low data rate and latency-tolerant transmissions. On the other hand, critical MTC represent those types of communications that require very low latency, ultra-high reliability, and high network availability. Therefore, they are also known as Ultra-Reliable Low-Latency Communications (URLLC). Supporting such type of MTC opens the door to many applications such as traffic safety, industry automation, emergency and disaster response, e-health services, and many other yet-to-appear applications.

Among the different wireless technologies, cellular networks are considered one of the most convenient technologies to provide the connectivity of critical MTC devices (MTCDs). This is by virtue of their advanced Radio Resource Management (RRM)

techniques and the availability of licensed spectrum that can guarantee the required stringent Quality of Service (QoS). Accordingly, the International Telecommunication Union (ITU) targets URLLC as a major use case, in addition to enhanced Mobile Broadband (eMBB) and massive MTC, in the requirements for the International Mobile Telecommunications 2020 and beyond (IMT-2020) [1]. The Third Generation Partnership Project (3GPP) is working on evolving the current Long-Term Evolution (LTE) standard, in addition to the New Radio (NR), to fulfill the Fifth-Generation (5G) requirements with backward compatibility [2]. Therefore, several enhancements in the PHYsical (PHY) and Medium Access Control (MAC) layers have been introduced in 3GPP Releases 14 and 15 to support critical MTC in LTE [3]. For instance, the concept of short transmission time intervals and supporting reduced processing time are considered in [4], in addition to fast uplink access on MAC in [5], as techniques to reduce the latency in LTE to serve critical MTC efficiently.

Massive Multiple-Input Multiple-Output (MIMO) is considered as a major technology to improve the spectral efficiency, processing complexity, and energy efficiency of LTE systems to fulfill the 5G requirements. Therefore, 3GPP targets employing tens of antennas at the eNodeB (eNB) to utilize the massive MIMO techniques [6]. These MIMO enhancements in LTE are standardized under the official name of Full-Dimension MIMO (FD-MIMO) [7]. In this case, the additional degrees of freedom introduced by massive MIMO can be exploited to serve critical MTC efficiently [8]. As analyzed in [9], the spatial degrees of freedom created by massive MIMO enable several beneficial properties for critical MTC such as high signal-to-noise ratio (SNR) links, spatial division multiplexing, and quasi-deterministic links that are immune to fast fading. In this regard, the study in [10] investigates the feasibility of the massive antenna systems to fulfill the stringent requirements of critical MTC in the uplink direction, testing different multi-antenna schemes such as coherent and non-coherent

receivers. On the other hand, the satisfaction of the requirements of critical MTC should be without sacrificing the QoS of the HTC traffic. This is due to the fact that the characteristics of critical MTC traffic is different than those of HTC in several aspects such as the data rate, the packet size, the latency-tolerance, and the reliability requirements. Therefore, and to achieve the goal of fulfilling the stringent QoS requirements of critical MTC without negative effects on HTC, RRM techniques should be optimized to serve both types of communications efficiently without degrading the system utility as well. Hence, in this chapter, we optimize the resource allocation and scheduling process for critical MTC, considering the coexistence of the HTC traffic, through exploiting massive MIMO techniques.

4.2.1 Related Work

Several recent studies consider the resource allocation problem of critical MTC without considering the coexistence of HTC traffic. In [11], the authors propose a downlink scheduler for reliable low latency users. First, they subdivide the users into two groups, high and low priority, according to the possibility of satisfying their QoS requirements in terms of maximum delay and packet error rate. Therefore, they serve the users who have QoS requirements that can be satisfied in the scheduling period first. However, they consider a special case of channel status feedback, in which a wideband report is used for the whole bandwidth. The study in [12] maximizes the energy efficiency in the downlink of Frequency Division Multiple Access (FDMA) systems that serve URLLC while considering their end-to-end delay and packet loss requirements. This is achieved by optimizing the transmit power, bandwidth and the number of active antennas. They adopt a finite blocklength analysis to approximate the achievable data rates of the users. Nevertheless, they do not consider Orthogonal FDMA (OFDMA)-based systems such as LTE. In [13], the study maximizes the

energy efficiency of URLLC in OFDMA-based radio access systems considering their QoS requirements of packet loss and latency. For this purpose, they optimize the packet dropping, power allocation, and bandwidth allocation policies. The authors in [14] extend the work in [12] and [13] by exploiting the multi-user diversity. However, they consider the downlink of FDMA-based cellular systems similar to [12]. In [15], the MTCDs are clustered based on their QoS characteristics, requirements and transmission protocols. Then, the aggregate data rate is maximized while considering the minimum data rate requirements of the devices. Nevertheless, separating the resource allocation processes for HTC and critical MTC, as discussed in the aforementioned works, does not optimize the overall resource allocation and reduces the gain resulting from multiuser diversity. Furthermore, this approach does not consider the impact of satisfying the stringent requirements of critical MTC on HTC traffic.

Therefore, studies consider the coexistence of MTC and HTC traffic types in the resource allocation problem. In [16], the authors consider splitting the radio resources between both types of users based on their buffer sizes. Then, every type of communication is scheduled separately. The resources are allocated fairly on the MTCDs considering their transmission deadlines. However, such splitting process of the radio resources before allocation does not optimize the allocation process at the system level. In [17, 18], the authors optimally maximize the aggregate data rate of the HTC traffic while considering the QoS requirements of all users. Nevertheless, they do not consider multiple antenna configurations that complicate the resource allocation process. This is due to the interference that can occur between users co-scheduled on the the same radio resources. Hence, the selection of the co-scheduled users should be taken into consideration while optimizing the resource allocation process. Moreover, they do not consider effective bandwidth and effective capacity concepts that can be used to provide statistical guarantees for the QoS of critical MTC as will be discussed

in Section 4.3.2.

4.2.2 Paper Contributions and Organization

The major contributions of this chapter can be summarized in the following:

- We formulate the resource allocation problem of critical MTC coexistent with HTC in massive MIMO LTE networks such that the system utility is maximized while satisfying the different QoS requirements of both types of communications. In this regard, we use the effective bandwidth [19] and effective capacity [20] concepts to design the resource allocation constraints from a cross-layer perspective to provide statistical guarantees for the QoS requirements of critical MTC in terms of probability of delay-bound violation. This considers both the PHY layer parameters and the buffer dynamics of the devices. Then, we formulate an equivalent instantaneous resource allocation problem exploiting the ergodicity of the service processes. However, an exponential computational complexity is required to calculate the global optimal solution of the formulated optimization problem that is NP-hard, as will be discussed in Section 4.3.

- Therefore, we propose a computationally-efficient algorithm for the formulated resource allocation problem that can be implemented in practice. For this purpose, we utilize the matching theory [21] to formulate the resource allocation problem as a matching process that can be solved efficiently with much lower computational complexity compared to that of the global optimal solution. In this regard, we analyze the computational complexity of the proposed algorithms in big-O notation and discuss and prove the convergence and stability of the proposed matching processes. In addition, the optimality of the proposed resource allocation scheme is investigated. Moreover, we run extensive simulations to

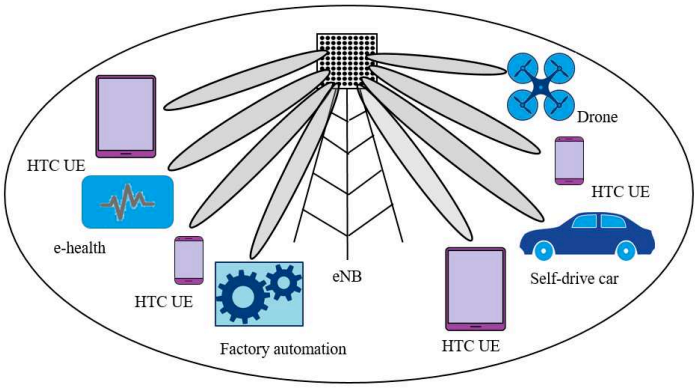


Fig. 4.1: An eNB with massive antennas serving critical MTCDs coexistent with HTC UEs in an LTE cell.

evaluate the performance of the proposed matching-based resource allocation technique and compare it with other algorithms from different scheduling techniques. The statistics of the major parameters impacting the computational complexity of the proposed algorithms are calculated.

The rest of the chapter is organized as follows. In Section 4.3, we discuss the adopted system model and formulate the resource allocation problem. The proposed matching-based resource allocation technique is presented in Section 4.4. Then, in Section 4.5, the proposed scheme is analyzed form the practical and computational perspective. The simulation results are presented and discussed in Section 4.6. Finally, the study is concluded in Section 4.7.

4.3 System Model and Problem Formulation

4.3.1 System Model and General Formulation

We consider the resource allocation and scheduling of the uplink transmissions of single-antenna users in a single LTE cell that is served by a single eNB, as shown in Fig. 4.1. Assume that the set of users is indexed by $\mathcal{U} = \mathcal{H} \cup \mathcal{M} = \{1, \cdots, u, \cdots, U\}$,

Table 4.1: Frequently Used Symbols and Notations of Chapter 4

Symbol	Description
$\mathcal{K},\mathcal{U},\mathcal{H},\mathcal{M}$	Sets of PRBs, users, HTC UEs, MTCDs, respectively
K, U, H, M	Cardinalities of $\mathcal{K},\mathcal{U},\mathcal{H},\mathcal{M}$, respectively
$\mathcal{K}_u$	Subset of PRBs assigned to user u
$\mathcal{C}_k$	Set of co-scheduled users on PRB k
A	Number of antennas at eNB
$\mathbf{y}_k$	Received signal on kth PRB
$\mathbf{h}_{u,k}$	Channel gain of uth user on kth PRB
$\gamma_{u,k}$	SNR of uth user on kth PRB
$P_{u,k}$	Transmit power of uth user on kth PRB
N_k	Power spectral density of AWGN on kth PRB
$\mathbf{v}_{u,k}$	Beamforming vector of uth user on kth PRB
T	Period of one TTI
i	TTI index
λ_u	Average arrival rate of uth user
$\mathcal{A}$	Arrival process
$\mathcal{S}$	Service process
$\bar{R}_u^{min}$	Minimum average rate of the uth user, $u \in \mathcal{H}$
K_u^{max}	Maximum no. of PRBs can be assigned to uth user
C_k^{max}	Maximum no. of users co-scheduled on kth PRB
D_u	Delay of uth user
D_u^{max}	Delay bound of uth user
ε_u	Maximum allowed PDBV of uth user
R_u	Achievable data rate of the uth user
$R_{u,k}$	Achievable data rate of the uth user on kth PRB
$x_{u,k}$	Indicator weather PRB k is assigned to user u or not
θ_u	QoS exponent of the uth user
Λ_u	Effective bandwidth of uth user
κ_u	Effective capacity of uth user
μ	Assignment operation of the matching process
$\varrho_{u,k}$	Desirability between user u and PRB k
Ψ	System utility function

where $\mathcal{H}$ is a set of HTC UEs and $\mathcal{M}$ is a set of critical MTCDs. Suppose that the number of HTC UEs and critical MTCDs in the cell are H and M, respectively. The system bandwidth is divided into Physical Resource Blocks (PRBs) of 180 KHz bandwidth that are indexed by $\mathcal{K} = \{1, \cdots, k, \cdots, K\}$. A user can use a PRB for uplink transmission for a time period known as the Transmission Time Interval (TTI). The frequently used symbols and notations are summarized in Table 4.1.

Assume that the eNB uses A antennas, where $A \gg U$. Such a massive number of antennas is deployed to utilize beamforming at the eNB for the uplink reception.

Therefore, a set, $\mathcal{C}_k$ of users can be co-scheduled on the same PRB k. That is, $\mathbf{y}_k \in \mathbb{C}^{A\times 1}$, which is the received signal vector at the eNB on the kth PRB, is calculated by

$$\mathbf{y}_k = \sum_{u\in\mathcal{C}_k} \mathbf{h}_{u,k}\sqrt{P_{u,k}}s_{u,k} + \mathbf{n}_k, \tag{4.1}$$

where $s_{u,k} \in \mathbb{C}$ is the data signal transmitted by the uth user on the kth PRB, which is normalized to unit power, $\mathbf{n}_k \in \mathbb{C}^{A\times 1}$ is the receiver AWGN noise vector on the kth PRB, which is a complex Gaussian vector with zero mean and covariance matrix of $N_k\mathbf{I}_A$, where $\mathbf{I}_A$ is the identity matrix of size A, and $P_{u,k}$ is the transmit power on the kth PRB by the uth user. The channel between the eNB and the uth user on the kth PRB is represented by $\mathbf{h}_{u,k} \in \mathbb{C}^{A\times 1}$ which is calculated by

$$\mathbf{h}_{u,k} = \sqrt{Z_u/L_u}\mathbf{f}_{u,k}, \tag{4.2}$$

where L_u is the power path loss, Z_u is the shadowing power gain, and $\mathbf{f}_{u,k}$ is the small-scale fading between the device and the eNB on the kth PRB, which is assumed to be independent and identically distributed complex Gaussian.

The received signal, $\mathbf{y}_k$, is multiplied by a unit-norm receive beamforming vector, $\mathbf{v}_{u,k} \in \mathbb{C}^{A\times 1}$, to spatially discriminate the signal sent by the uth user on the kth PRB from the interfering signals of other co-scheduled users on the same PRB, $\{u' \neq u : \ u' \in \mathcal{C}_k\}$. Therefore, the uplink SINR of the signal from the uth user on the kth PRB can be calculated by [22]

$$\gamma_{u,k} = \frac{\frac{P_{u,k}}{N_k}|\mathbf{h}_{u,k}^H\mathbf{v}_{u,k}|^2}{\sum_{\forall u'\neq u, u'\in\mathcal{C}_k}\frac{P_{u',k}}{N_k}|\mathbf{h}_{u',k}^H\mathbf{v}_{u,k}|^2 + \mathbf{v}_{u,k}^H\mathbf{I}_A\mathbf{v}_{u,k}}. \tag{4.3}$$

Consequently, the maximum achievable data rate of user u over PRB k is

$$R_{u,k} = B \log_2(1 + \gamma_{u,k}), \tag{4.4}$$

where $B = 180$ KHz, is the bandwidth of one PRB.

Every TTI, the scheduler in the eNB assigns the PRBs to the users such that the system utility is maximized while satisfying the QoS requirements of the users in the cell. According to [1], achieving high data rates for critical MTC is of low importance since their transmissions are characterized by their low data rate [23] and small packet size [24]. However, satisfying their latency and reliability requirements is crucial. On the other hand, the QoS of HTC improves by increasing their data rates. Therefore, maximizing the data rate of all users in the cell impacts the resource utilization negatively. This is because maximizing the data rate of critical MTC does not improve their QoS, given that their latency requirements are satisfied. Nevertheless, this data rate is at the expense of that of the HTC UEs.

As a consequence, we formulate the resource allocation problem such that the aggregate data rate of the HTC traffic is maximized while considering the QoS requirements of all users as constraints. That is, the optimization problem of the resource allocation process is formulated as follows:

$$\max_{\{\mathcal{K}_1,\cdots,\mathcal{K}_U\}\in\mathcal{K}} \quad \sum_{u\in\mathcal{H}} R_u \tag{4.5}$$

$$\text{s.t.} \quad \mathbb{E}\{R_u\} \geq \bar{R}_u^{min}, \ \forall u \in \mathcal{H} \tag{4.5a}$$

$$\Pr[D_u \geq D_u^{max}] \leq \varepsilon_u, \ \forall u \in \mathcal{M} \tag{4.5b}$$

$$|\mathcal{K}_u| \leq K_u^{max}, \ \forall u \in \mathcal{M} \tag{4.5c}$$

$$|\mathcal{C}_k| \leq C_k^{max}, \ \forall k \in \mathcal{K}, \tag{4.5d}$$

where $\mathcal{K}_u \in \mathcal{K}$ is the subset of PRBs assigned to the uth user, R_u is the maximum achievable data rate of user u over the subset of PRBs assigned to it, and $\mathbb{E}\{R_u\}$ is its average rate. To guarantee a minimum average rate for each HTC user, constraint (4.5a) is used, where $\bar{R}_u^{min}$ is the required minimum average rate of user u. On the other hand, we use accurate statistical guarantees for the latency requirements of critical MTC. For this purpose, we ensure that the probability of delay bound violation (PDBV) of each critical MTCD is under a certain threshold ε_u as in constraint (4.5b), where D_u^{max} is the delay bound for the uth MTCD. Therefore, given that the packets that miss their deadlines are dropped, the parameter ε_u represents one component of the reliability guarantees of MTCD u. Constraint (4.5c) is used to ensure a maximum number of allowed PRBs to be assigned to MTCDs. For example, in LTE Release 13, the number of PRBs that are assigned to MTCDs is limited to 6. Constraint (4.5d) is expressed to limit the number of co-scheduled users on PRBs as used in the framework of users pairing as in [25], for instance. In (4.5c) and (4.5d), K_u^{max} is the maximum number of PRBs that can be assigned to MTCD u and C_k^{max} is the maximum number of co-scheduled users allowed on PRB k. As discussed in [26], non-contiguous resource allocations are allowed in the uplink of LTE-Advanced. This enhances the spectral efficiency as discussed in [27] thanks to using frequency-selective scheduling.

4.3.2 Cross-Layer Design and Formulation

To provide statistical guarantees for the satisfaction of the latency requirements of critical MTC, a cross-layer design is required to consider their buffer dynamics as well as PHY layer parameters. For this purpose, we use the effective bandwidth and effective capacity concepts.

The resource allocation and scheduling process determines the data rate of every user in every TTI and controls the dynamics of the queues of the devices, as shown

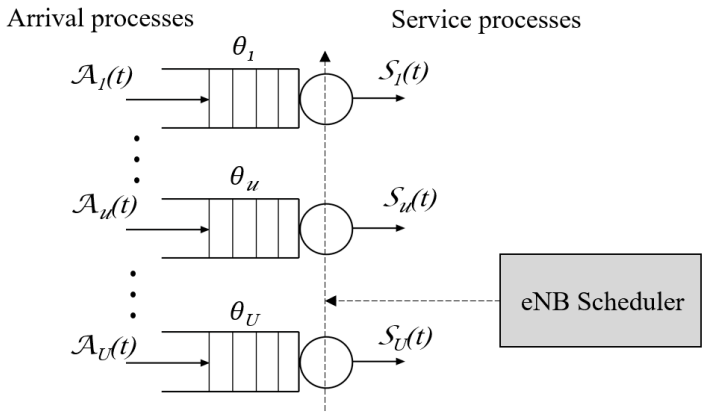


Fig. 4.2: A cross-layer perspective of the eNB scheduler.

in Fig. 4.2. Let us define the arrival and service processes, in bits, of user u as $\mathcal{A}_u(t)$ and $\mathcal{S}_u(t)$, respectively. According to the large deviations theory, the PDBV of the queue can accurately be approximated by [19]:

$$\Pr[D_u(t) \geq D_u^{max}] \approx e^{-\theta_u \delta_u D_u^{max}}, \tag{4.6}$$

where δ_u depends on both the arrival and service processes as will be discussed below and θ_u is known as the QoS exponent that characterizes the queue length decaying rate where a smaller θ_u represents a looser QoS constraint and vice versa.

The effective bandwidth [19] of the arrival process of user u is defined as the minimum constant service rate that can serve that process with a guaranteed QoS exponent θ_u such that

$$\Pr[D_u(t) \geq D_u^{max}] \approx e^{-\theta_u \delta_u D_u^{max}} \leq \varepsilon_u, \tag{4.7}$$

and is calculated by

$$\Lambda_u(\theta_u) = \lim_{t \to \infty} \frac{1}{t\theta_u} \ln \mathbb{E}\{e^{\theta_u \mathcal{A}_u(t)}\}. \tag{4.8}$$

In a similar manner, the effective capacity [20] of the service process of the uth user is defined as the maximum constant arrival rate that can be served by the process with a guaranteed QoS exponent θ_u, and is calculated by

$$\kappa_u(\theta_u) = -\lim_{t\to\infty} \frac{1}{t\theta_u} \ln \mathbb{E}\{e^{-\theta_u \mathcal{S}_u(t)}\}. \tag{4.9}$$

Therefore, the effective capacity of a wireless channel converges to the ergodic capacity when the QoS constraints are relaxed as discussed in [28].

The parameter δ_u can be calculated by deriving the rate at which the effective capacity and effective bandwidth curves intersect [29]. That is, $\delta_u = \kappa_u(\theta_u^*) = \Lambda(\theta_u^*)$. For instance, for a Poisson process, the parameter δ_u can be calculated as follows [30]:

$$\delta_u = \lambda_u \left(\frac{e^{\theta_u^*} - 1}{\theta_u^*} \right), \tag{4.10}$$

where λ_u is the arrival rate of the Poisson process.

Accordingly, to guarantee a certain QoS exponent for an MTCD u, the effective capacity of the service process should satisfy the following inequality

$$\kappa_u(\theta_u) \geq \Lambda_u(\theta_u), \tag{4.11}$$

where the guaranteed QoS exponent, θ_u, represents the required QoS level $(D_u^{max}, \varepsilon_u)$ and can be derived from (4.7) as

$$\theta_u = \frac{-\ln \varepsilon_u}{\delta_u D_u^{max}}. \tag{4.12}$$

To derive the effective capacity of the service process that represents the serviced bits at time t, we assume that the data rate of user u at the ith TTI is $R_u[i]$. Therefore,

the sequence $\{R_u[i]T : i = 1, 2, 3, \cdots\}$, where T is the TTI period, is a discrete-time stationary and ergodic random process. Hence, the service process for the uth user is

$$\mathcal{S}_u[t] = \sum_{i=1}^{t} R_u[i]T. \tag{4.13}$$

Due to the fact that the sequence $\{R_u[i]T : i = 1, 2, 3, \cdots\}$ is uncorrelated, the effective capacity of the uth user in (4.9) reduces to [31]:

$$\kappa_u(\theta_u) = \frac{-1}{\theta_u} \ln \mathbb{E}\{e^{-\theta_u R_u[i]T}\}. \tag{4.14}$$

From the previous discussion, the PDBV constraint of critical MTCDs in (4.5b) can be expressed in a cross-layer perspective using (4.11) and (4.14). That is, the equivalent optimization problem to that in (4.5) is

$$\max_{\mathbf{X}} \quad \sum_{u\in\mathcal{H}} \sum_{k=1}^{K} R_{u,k} x_{u,k} \tag{4.15}$$

$$\text{s.t.} \quad \mathbb{E}\{R_u\} \geq \bar{R}_u^{min}, \ \forall u \in \mathcal{H} \tag{4.15a}$$

$$\frac{-1}{\theta_u} \ln \mathbb{E}\{e^{-\theta_u R_u[i]T}\} \geq \Lambda_u, \ \forall u \in \mathcal{M} \tag{4.15b}$$

$$\sum_{k=1}^{K} x_{u,k} \leq K_u^{max}, \ \forall u \in \mathcal{M} \tag{4.15c}$$

$$\sum_{u=1}^{U} x_{u,k} \leq C_k^{max}, \ \forall k \in \mathcal{K} \tag{4.15d}$$

$$x_{u,k} \in \{0, 1\}, \ \forall u \in \mathcal{U}, \ k \in \mathcal{K}, \tag{4.15e}$$

where $\mathbf{X}$ is a $U \times K$ binary indicator matrix such that $x_{u,k}$ indicates whether PRB k is assigned to user u. Constraints (4.15a)–(4.15d) are equivalent to (4.5a)–(4.5d), respectively. Constraint (4.15e) is used to restrict $x_{u,k}$ to binary values.

The optimization problem in (4.15) falls in the Binary Nonlinear Programming

(BNLP) category. This type of problems can be optimally solved using exhaustive search or algorithms such as the Branch and Bound (BB). However, the computational complexity of such algorithms is exponential which makes the problem NP-hard [32]. Therefore, these algorithms cannot be used in real-time processing such as in resource allocation and scheduling. Therefore, we propose computationally-efficient algorithms as a trade-off between the complexity and the performance so that they can be used in practice as resource allocation and scheduling schemes.

4.4 Matching-Based Resource Allocation

In this section, we formulate an instantaneous resource allocation problem that can be solved every TTI such that the long-term constraints (4.15a) and (4.15b) are satisfied. Then, utilizing the matching theory, we formulate the instantaneous problem as a two-sided matching process. Finally, we propose a complete matching-based resource allocation algorithm.

4.4.1 Formulation of the Instantaneous Resource Allocation Problem

Theorem 4.4.1 can be used to restrict the instantaneous data rates of the users such that their average data rate or PDBV constraints be satisfied in the long-term. That is, we derive data rate constraints equivalent to the constraints in (4.15a) and (4.15b) as follows.

Theorem 4.4.1. *The long-term constraints in (4.15a) and (4.15b) for the HTC and critical MTC, respectively, can be fulfilled if the following necessary and sufficient set*

of constraints is satisfied:

$$R_u[i] \geq R_u^{min}[i], \ \forall u \in \mathcal{U}, \tag{4.16}$$

where, $R_u^{min}[i]$ is the instantaneous minimum data rate at the ith TTI for the uth user to fulfill its long-term constraint and is calculated by (4.17) as

$$R_u^{min}[i] = \begin{cases} i\bar{R}_u^{min} - (i-1)R_u^{avg}[i-1], & \forall u \in \mathcal{H} \\ \frac{1}{\theta_u T} \ln\left(ie^{-\theta_u \Lambda_u} - (i-1)\Phi_u^{avg}[i-1]\right), & \forall u \in \mathcal{M} \end{cases}, \tag{4.17}$$

where

$$R_u^{avg}[i] = \begin{cases} \frac{R_u[i]+(i-1)R_u^{avg}[i-1]}{i}, & i \geq 2 \\ R_u[i], & i = 1 \end{cases}, \tag{4.18}$$

$$\Phi_u^{avg}[i] = \begin{cases} \frac{\Phi_u[i]+(i-1)\Phi_u^{avg}[i-1]}{i}, & i \geq 2 \\ \Phi_u[i], & i = 1 \end{cases}, \tag{4.19}$$

$$\Phi_u[i] = e^{-\theta_u R_u[i] T}. \tag{4.20}$$

Proof. To derive the minimum instantaneous rate of the set of HTC UEs, define $R_u^{avg}[i]$ as the cumulative moving average (CMA) of $R_u[i]$ at the ith TTI. This can be calculated using (4.18). This CMA represents the estimation of $\mathbb{E}\{R\}$, at the ith TTI, due to the ergodicity of the random process composed by the sequence $\{R_u[i] : \ i = 1, 2, 3, \cdots\}$. Therefore, the constraint in (4.15a) can be satisfied by fulfilling the following instantaneous constraint

$$R_u^{avg}[i] \geq \bar{R}_u^{min}, \ \forall u \in \mathcal{H}. \tag{4.21}$$

Using (4.18), we can write (4.21) as

$$R_u[i] \geq i\bar{R}_u^{min} - (i-1)R_u^{avg}[i-1], \ \forall u \in \mathcal{H}. \tag{4.22}$$

Therefore, the equivalent minimum instantaneous rate for the HTC UEs can be given by (4.17).

Similarly, to derive the minimum instantaneous data rate of the MTCDs, we define $\Phi_u[i]$ as in (4.20) and $\Phi_u^{avg}[i]$ as the CMA of $\Phi_u[i]$ at the ith TTI as given in (4.19). Similarly, this represents the estimation of $\mathbb{E}\{e^{-\theta_u R_u[i]T}\}$ since the random process $\{R_u[i]T: \ i = 1, 2, 3, \cdots\}$ is ergodic. Thus, the constraint in (4.15b) can be expressed as

$$\Phi_u^{avg}[i] \leq e^{-\theta_u \Lambda_u}. \tag{4.23}$$

Using (4.19), (4.23) can be rewritten in the following form

$$e^{-\theta_u R_u[i]T} \leq ie^{-\theta_u \Lambda_u} - (i-1)\Phi_u^{avg}[i-1]. \tag{4.24}$$

The last inequality can be written as in the form used in (4.16). Therefore, the minimum instantaneous data rate of the critical MTCDs can be derived as in (4.17). □

Using the equivalent set of constraints as in (4.16) in place of (4.15a) and (4.15b), we can derive an instantaneous resource allocation problem that is equivalent to (4.15)

at the ith TTI as follows

$$\max_{\mathbf{X}} \quad \sum_{u\in\mathcal{H}}\sum_{k=1}^{K} R_{u,k}[i]x_{u,k} \tag{4.25}$$

$$\text{s.t.} \quad R_u[i] \geq R_u^{min}[i],\ \forall u \in \mathcal{U} \tag{4.25a}$$

$$\sum_{k=1}^{K} x_{u,k} \leq K_u^{max},\ \forall u \in \mathcal{M} \tag{4.25b}$$

$$\sum_{u=1}^{U} x_{u,k} \leq C_k^{max},\ \forall k \in \mathcal{K} \tag{4.25c}$$

$$x_{u,k} \in \{0,1\},\ \forall u \in \mathcal{U},\ k \in \mathcal{K}. \tag{4.25d}$$

To solve the equivalent instantaneous problem in (4.25), we utilize the matching theory to devise a computationally-efficient algorithm.

4.4.2 Matching Model and Formulation

To formulate the resource allocation problem in (4.25) as a centralized matching process, we assume that $\mathcal{U}$ and $\mathcal{K}$ are two disjoint sets of agents that are willing to maximize their utilities and satisfy their minimum requirements. After the PRB assignment process is complete, we say that (u,k) is a matched pair if PRB k is assigned to user u. Therefore, a two-sided matching μ for the considered resource allocation problem in (4.25) can be defined as follows.

Definition 4.4.2. *A matching μ that is equivalent to the resource allocation problem in (4.25) is defined as a mapping from the set $\mathcal{U}\cup\mathcal{K}$ into the set $\mathcal{U}\cup\mathcal{K}$ such that for any $u\in\mathcal{U}$ and $k\in\mathcal{K}$:*

(i) $\mu(u)\subseteq\mathcal{K}$,

(ii) $\mu(k)\subseteq\mathcal{U}$,

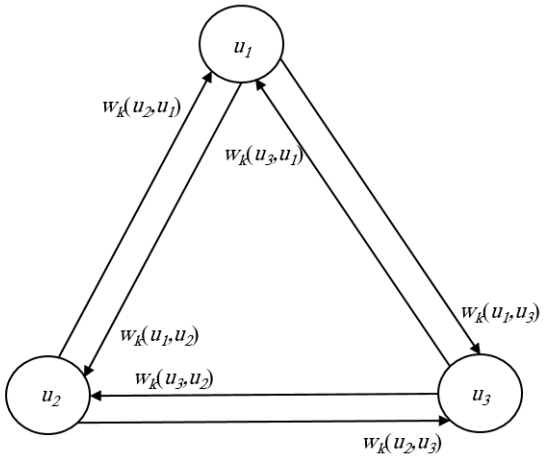


Fig. 4.3: Weighted directed friendship network between users.

(iii) $|\mu(k)| \leq C_k^{max},\ \forall k \in \mathcal{K}$,

(iv) $|\mu(u)| \geq q_u^{min},\ \forall u \in \mathcal{H}$,

(v) $q_u^{min} \leq |\mu(u)| \geq q_u^{max},\ \forall u \in \mathcal{M}$,

(vi) $k \in \mu(u) \Longleftrightarrow u \in \mu(k)$.

Condition (i) indicates that every user $u \in \mathcal{U}$ can be matched to a set of PRBs. Also, every PRB $k \in \mathcal{K}$ can be matched to a set of users as indicated in condition (ii). Therefore, this matching process falls in the many-to-many matching category. Condition (iii) represents the maximum number of co-scheduled users per PRB k. Condition (iv) represents the minimum rate requirement of the HTC UEs, where the minimum quota, q_u^{min}, is the cardinality of the set of PRBs that satisfy this constraint. Similarly, condition (v) is formulated for the minimum rate and the maximum number of PRBs constraints of the MTCDs. Condition (vi) indicates that if a PRB k is matched to a user u, then it should be in its matched set of PRBs as well.

Due to the interference between users, the matching of every user u to every PRB k does not depend only on its channel conditions on this PRB. That is, every user

u cares about other users that are matched to the same PRBs. Therefore, similar to [33], we use a weighted, directed social network graph to model the relationship of every user to other users on every PRB to represent the interference between them as follows.

Definition 4.4.3. *The friendship network among users on every PRB k is modeled as a weighted graph $G = (\mathcal{N}, \Xi_k, w_k)$, where $\mathcal{N} = \mathcal{U}$ is the set of nodes, Ξ_k is the set of arcs between them on PRB k, and w_k are the weights that represent the relationship between users on the kth PRB, as shown in Fig. 4.3. The relationship between user u and u' on PRB k is weighted by*

$$w_k(u, u') = \frac{P_{u',k}}{N_k} |\boldsymbol{h}_{u',k}^H \boldsymbol{v}_{u,k}|^2. \tag{4.26}$$

To define the utility of agents, we first define the desirability between user u and PRB k, $\varrho_{u,k}$, as follows

$$\varrho_{u,k} = \frac{P_{u,k}}{N_k} |\mathbf{h}_{u,k}^H \mathbf{v}_{u,k}|^2. \tag{4.27}$$

Therefore, the utility of user u on PRB k depends on the desirability of user u and PRB k, and the weight of the relationship between u and other users co-scheduled on the same PRB, $\{u' \neq u : \ u' \in \mathcal{C}_k\}$. That is, the utility of user u on PRB k, $\Psi_{u,k}$, can be calculated by

$$\Psi_{u,k} = B \log_2 \left(1 + \frac{\varrho_{u,k}}{\sum_{\forall u' \neq u, u' \in \mathcal{C}_k} w_k(u, u') + \mathbf{v}_{u,k}^H \mathbf{I}_A \mathbf{v}_{u,k}} \right). \tag{4.28}$$

On the other hand, the utility of every PRB depends on the utilities of the HTC UEs

scheduled on this PRB. Therefore, the utility of PRB k can be calculated as

$$\Psi_k = \sum_{u \in \mu(k) \cap \mathcal{H}} \Psi_{u,k}. \tag{4.29}$$

Accordingly, to maximize the aggregate data rate of the HTC users, the matching assignment μ should maximize the system utility Ψ that is defined as follows

$$\Psi = \sum_{k=1}^{K} \Psi_k, \tag{4.30}$$

subject to the conditions in Definition 4.4.2

4.4.3 Matching-Based Resource Allocation Algorithm

We now propose the resource allocation algorithm that is based on the matching process formulated in Section 4.4.2. The matching process is a many-to-many assignment. However, two major challenges arise in this matching process. The first one is the lower quota bounds that are used for the minimum rate constraints. The second challenge is the externalities in the problem since the allocation of the PRBs to a certain user affects the other users that are co-scheduled on the same PRBs. To address these challenges, we perform the matching process in two phases, where each phase addresses one of the challenges. Algorithm 4.1 summarizes the proposed approach for solving the resource allocation problem in (4.15) and how the two phases of matching can be used to overcome the difficulty of the matching process.

In Algorithm 4.1, in every TTI, we construct the instantaneous resource allocation problem as in (4.25) and then derive a matching that solves it, as discussed in Section 4.4.2. To establish the matching process, every agent $u \in \mathcal{U}$, or $k \in \mathcal{K}$, composes its preference list $\mathcal{P}(u)$, or $\mathcal{P}(k)$, respectively, in which the agents in the opposite set are

Algorithm 4.1 Proposed Matching-Based Scheduling Algorithm

1: **for all** TTIs **do**
2: Construct the instantaneous equivalent resource allocation problem as in (4.25) by calculating the minimum instantaneous data rate required for all $u \in \mathcal{U}$ at current TTI using (4.17).
3: Formulate the instantaneous problem as a two-sided matching process by constructing the preference lists of all $u \in \mathcal{U}$ over $k \in \mathcal{K}$ and all $k \in \mathcal{K}$ over $u \in \mathcal{U}$ according to $\varrho_{u,k}$.
Matching Phase 1
4: Use Algorithm 4.2 to match the agents so that their minimum rate constraints are satisfied.
Matching Phase 2
5: Use Algorithm 4.3 to match the agents to maximize the data rate of the HTC users.
6: **end for**

ordered. Therefore, we say that PRB k is preferred to k' by user u which is expressed as $k \succ_u k'$, if k precedes k' in u's preference list, $\mathcal{P}(u)$. Similarly, if user u precedes u' in k's preference list $\mathcal{P}(k)$, we say that $u \succ_k u'$. The ordering of the agents of the opposite set depends on the desirability between the two agents, $\varrho_{u,k}$, as in (4.27). That is,

$$k \succ_u k' \Longleftrightarrow \varrho_{u,k} > \varrho_{u,k'}, \tag{4.31}$$

$$u \succ_k u' \Longleftrightarrow \varrho_{u,k} > \varrho_{u',k}. \tag{4.32}$$

The preferences of the agents are transitive. That is, if $u \succ_k u'$ and $u' \succ_k u''$, then $u \succ_k u''$.

In Phase 1 of the matching, the users are matched such that their minimum instantaneous rate requirements are satisfied without considering the maximization of the aggregate data rate of the HTC UEs. For this purpose, we use Algorithm 4.2 that is based in principle on the one-to-many Gale-Shapley algorithm [34] after adapting it to the many-to-many problem and considering the externalities and lower quota

Algorithm 4.2 Satisfy Minimum Rate Constraints

Set the status of all $u \in \mathcal{U}$ that have minimum rate constraints to 1 and others to 0, where a status of 1 indicates that the user is willing to propose and otherwise is status 0.
while any user's status is 1 **do**
  Listen to the first user willing to propose. Assume it is u^* and its preferred PRB is k^* which is not in current $\mu(u^*)$.
  if $|\mu(k^*)| < C_k^{max}$ **then**
    Match u^* to k^* by updating their match lists $\mu(u^*)$ and $\mu(k^*)$ respectively.
    for all $u \in \mu(k^*)$ **do**
      Update its rate on the current PRB k^* considering the interference of other co-scheduled users.
      if $R_u < R_u^{min}[i]$ and $|\mu(u)| < K_u^{max}$ **then**
        Set the status of u to 1.
      else
        Set the status of u to 1.
      end if
    end for
  else
    Let k^* select the preferred set of users from the matched set and the candidate one u^* according to its preference list.
    if the rejected user is the proposing one u^* **then**
      Remove k^* from the preference list of u^* and clear its status if its preference list became empty.
    else
      Update the rate and status of the accepted set.
      Update the preference list, rate, and status of the rejected user.
    end if
  end if
end while
Feasibility test
if any $u \in \mathcal{U}$ still not satisfied **then**
  Problem is infeasible.
end if

bounds. Then, in Phase 2, the aggregate data rate of the HTC users is maximized by utilizing Algorithm 4.3. In Algorithm 4.3, the users are added, deleted, and swapped such that the system utility Ψ is maximized without violating the minimum rate constraints that were satisfied in Phase 1. These three operations are similar to the swap-matching techniques studied for one-to-many problems in [33] to overcome the

Algorithm 4.3 Maximize the HTC Data Rate

1: **while** There is still approved addition/deletion/swap **do**
Step 1: Add users to improve the utility of PRBs
2: **for all** $k \in \mathcal{K}$ **do**
3: Determine the unmatched users and sort them according to the preference list of the PRB k, $\mathcal{P}(k)$.
4: Consider the users in this candidate list in order, to be added to the current match $\mu(k)$. The approved user must yield a better utility of the PRB, Ψ_k, without violating the minimum rate requirements of the currently matched users, $\mu(k)$, in addition to the other conditions in Definition 4.4.2.
5: Add the approved users to the current match.
6: Update the rate of the new matched set of users.
7: Update the utility of PRB k, Ψ_k.
8: **end for**
Step 2: Delete users to improve the utility of PRBs
9: **for all** $k \in \mathcal{K}$ **do**
10: Search in the matched users, $\mu(k)$, for the ones that can be unmatched to PRB k such that the utility function of the PRB, Ψ_k, would improve without violating their minimum rate requirements.
11: Unmatch the approved users from PRB k.
12: Update the rate of the matched and rejected users on PRB k.
13: Update the utility of PRB k, Ψ_k.
14: **end for**
Step 3: Swap users to improve the system utility
15: **for all** $u \in \mathcal{U}$ **do**
16: Search $\mathcal{U} \backslash \{u\}$ for an approved swap that can improve the system utility function, Ψ, without violating the minimum rate constraints of the users.
17: Implement the approved swaps.
18: Update the rate of the affected users.
19: Update the utilities of affected PRBs.
20: **end for**
21: **end while**

externalities in the problem. However, we use addition and deletion operations in addition to the swap operation in our many-to-many problem. Also, we consider the lower quota bounds in all operations.

4.5 Analysis of the Proposed Methods

In this section, we analyze the performance of the proposed resource allocation scheme from a practical perspective. For this purpose, we analyze the stability and convergence of the proposed matching algorithms. In addition, we discuss the optimality and computational complexity of the proposed scheme.

4.5.1 Stability

The stability of the proposed resource allocation scheme in Algorithm 4.1 depends on that of the matching phase in Algorithm 4.3. This is because the other matching phase in Algorithm 4.2 is used mainly to satisfy the minimum instantaneous rate requirements of the users. To define the stability of Algorithm 4.3, we first define the swap, addition, and deletion matchings as follows.

Definition 4.5.1. *Swap, $\mu^{s}_{u_1,u_2}$, addition, $\mu^{a}_{u,k}$, and deletion, $\mu^{d}_{u,k}$, matchings are defined respectively as follows:*

- $\mu^{s}_{u_1,u_2} = \{\mu \backslash \{(u_1,k_1),(u_2,k_2)\} \cup \{(u_1,k_2),(u_2,k_1)\}\}$,
  $k_1 \in \mu(u_1),\ k_2 \in \mu(u_2)$,
- $\mu^{a}_{u,k} = \mu \cup (u,k)$, *and*
- $\mu^{d}_{u,k} = \mu \backslash (u,k)$.

Given the definition of swap, addition, and deletion matchings, the stability of the matching scheme in Algorithm 4.3 can be defined as follows.

Definition 4.5.2. *A matching μ is stable if and only if there are no u, u', k such that*

1. $\Psi(\mu^{s}_{u,u'}) > \Psi(\mu)$,

2. $\Psi(\mu^a_{u,k}) > \Psi(\mu)$, *or*

3. $\Psi(\mu^d_{u,k}) > \Psi(\mu)$.

This is given that the matchings $\mu^s_{u,u'}$, $\mu^a_{u,k}$, *and* $\mu^d_{u,k}$ *satisfy the minimum instantaneous rate requirements of all users in addition to the other conditions in Definition 4.4.2.*

The stability of Algorithm 4.3 is analyzed as follows.

Lemma 4.5.3. *If the matching scheme in Algorithm 4.3 converges to a matching* μ^*. *Then, this matching* μ^* *is stable as defined in Definition 4.5.2.*

Proof. Assume that there are u', u'', k' that can yield $\Psi(\mu^a_{u',k'}) > \Psi(\mu)$, $\Psi(\mu^d_{u',k'}) > \Psi(\mu)$, or $\Psi(\mu^s_{u',u''}) > \Psi(\mu)$, and the new matchings satisfy the conditions in Definition 4.4.2. Then, this new matching would be approved in Step 1, 2, or 3, respectively, in Algorithm 4.3. This is because Steps 1, 2, and 3 in Algorithm 4.3 search for all approved addition, deletion, and swap operations, respectively, which improves the system utility Ψ without violating the conditions in Definition 4.4.2. Accordingly, these u, u', k cannot exist given that the algorithm converged to a matching μ^*. Consequently, the matching μ^* is stable. □

4.5.2 Convergence

The convergence of the proposed resource allocation scheme depends on that of the matching algorithms in Algorithm 4.2 and Algorithm 4.3. Therefore, in Theorem 4.5.4 we discuss the convergence of Algorithms 2 and 3 as follows.

Theorem 4.5.4. *The proposed matching schemes in Algorithm 4.2 and Algorithm 4.3 converge after a finite number of iterations.*

Proof. In Algorithm 4.2, every user u proposes to its preferred PRB in its preference list, $\mathcal{P}(u)$, in order. If it is rejected by a PRB, it deletes it from its preference list and proposes to the next one until it satisfies its requirements, or its preference list becomes empty. Since the number of PRBs is limited, the preference list of every user u is limited as well. Therefore, the number of proposals, and hence iterations, is limited. Consequently, Algorithm 4.2 converges after a finite number of iterations.

In Algorithm 4.3, after every approved addition, deletion, or swap operation, the new matching improves the system utility. That is, if the matching after every approved operation is as follows

$$\mu^{(1)},\ \mu^{(2)},\ \cdots, \mu^{(j-1)},\ \mu^{(j)},\ \cdots, \mu^{(final)}, \tag{4.33}$$

then $\Psi(\mu^{(j)}) > \Psi(\mu^{(j-1)})$. In other words, the system utility improves from every matching to the next. Due to the limited number of users and PRBs, the number of matchings is finite. In addition, the sum rate of the HTC UEs, which is the system utility, $\Psi(\mu)$, has an upper bound. Therefore, there is a round in which there is no further operation can be approved by the algorithm. Consequently, Algorithm 4.3 converges after a finite number of approved operations. □

4.5.3 Optimality

To analyze the optimality of the proposed resource allocation technique in Algorithm 4.1, we investigate how Algorithms 2 and 3 are used to get to a final solution for the problem in (4.25). As previously discussed, Algorithm 4.2 is mainly used to find a feasible solution that satisfies the minimum instantaneous rate requirements in addition to the remaining constraints in (4.25). However, Algorithm 4.3 is used to maximize the aggregate data rate of the HTC users, which is the objective function

of (4.25), without violating the feasibility of the solution. Hence, the optimality of Algorithm 4.1 depends on that of Algorithm 4.3.

To analyze the optimality of Algorithm 4.3, we first discuss the relationship between the local maxima of the problem in (4.25) and the stability of the solution as a matching scheme as follows.

Theorem 4.5.5. *All local maxima of the objective function of the problem in (4.25) represent a stable matching as defined in Definition 4.5.2.*

Proof. Assume that a resource allocation pattern, that is represented by the matching μ^*, is a local maximum to the optimization problem in (4.25). If μ^* is not a stable matching, then, according to Definition 4.5.2, there is at least one addition, deletion, or swap operation that can yield a better matching that has a better system utility function $\Psi(\mu)$. Since the system utility function Ψ is the same as the objective function of the problem in (4.25), this contradicts the assumption that μ^* is a local maximum. Therefore, μ^* must be a stable matching. □

Consequently, the optimality of Algorithm 4.1 can be proved as in the following lemma.

Lemma 4.5.6. *The matching-based resource allocation scheme in Algorithm 4.1 yields a local optimal solution for the optimization problem in (4.25).*

Proof. This is a direct the result of Theorem 4.5.5 and the stability proof of Algorithm 4.1 that is based on Lemma 4.5.3. □

4.5.4 Computational Complexity

To analyze the computational complexity of the proposed resource allocation scheme in Algorithm 4.1, we calculate the worst case computational complexity of every step

in Algorithm 4.1 in terms of big-O notation. For this purpose, we first analyze the computational complexity of the steps of Algorithm 4.2 and 4.3.

The worst case computational complexity of the steps of Algorithm 4.2 can be summarized as follows:

- step 1 requires $\mathcal{O}(U)$,
- steps 2-23 require $\mathcal{O}(UKC_k^{max})$, and
- steps 24-26 require $\mathcal{O}(U)$.

The steps of Algorithm 4.3 have the following computational complexity:

- steps 2-8 require $\mathcal{O}(K(min(H, C_k^{max}))^2)$,
- steps 9-14 require $\mathcal{O}(KC_k^{max})$, and
- steps 15-20 require $\mathcal{O}(U(U-1)C_k^{max}K_u^{max})$.

Accordingly, we can analyze the worst case computational complexity of every step of Algorithm 4.1 as follows:

- step 2 requires $\mathcal{O}(U)$,
- step 3 requires $\mathcal{O}(UK^2) + \mathcal{O}(KU^2)$,
- step 4 (Algorithm 4.2) requires $\mathcal{O}(KUC_k^{max})$, and
- step 5 (Algorithm 4.3) requires $\mathcal{O}(r(U^2K_u^{max}C_k^{max}))$,

where r is the number of rounds implemented in Algorithm 4.3. Numerical evaluations for this parameter are presented in Section 4.6.

Therefore, step 5 dominates the total complexity of the proposed resource allocation scheme. In fact, this is the computational complexity of the swap operations

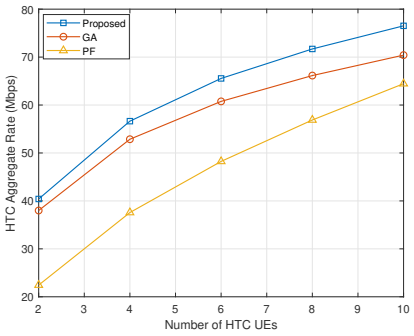
Table 4.2: Simulation Parameters of Chapter 4

Parameter	**Value**
Cell radius (C)	500 m
Number of eNBs	1
Simulation time	200 TTI
Number of runs	10
Path loss ($PL0 + 10n \log_{10} d$)	$128.1 + 37.6 \log_{10}(d)$, d in km [35]
Standard deviation of shadowing (σ)	8 dB
Transmitter power (P)	15 dBm
Power spectral density of noise	-174 dBm/Hz
Noise figure	18 dB
Number of PRBs (K)	25
Distribution of MTCDs/UEs	Fixed and uniform
HTC arrival rate ($\lambda_u,\ u \in \mathcal{H}$)	$64, 128, 192, 256$ kbps
MTC arrival rate ($\lambda_u,\ u \in \mathcal{M}$)	$10, 20, 30, 40$ kbps
Arrival rates distribution	Uniform
Delay bound ($D_u^{max},\ u \in \mathcal{M}$)	0.2 ms
Maximum PDBV ($\varepsilon_u,\ u \in \mathcal{M}$)	10^{-2}
$\bar{R}_u^{min},\ u \in \mathcal{H}$	λ_u

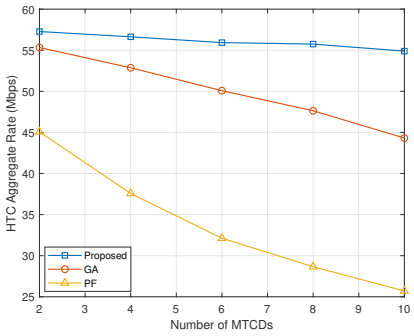
that are used to maximize the aggregate data rate of the HTC users. However, the complexity of the algorithm is still much lower than that of the global optimal solution. This is because, as mentioned above, the computational complexity of the global optimal solution of BNLP problem is exponential, which makes the problem NP-hard [32].

4.6 Experimental Results

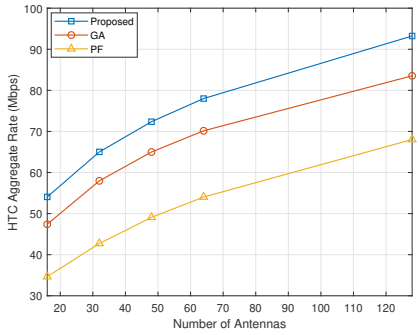
In this section, we present and discuss the results of the simulation experiments performed to evaluate the performance of the proposed matching-based resource allocation scheme. We compare the performance of the proposed algorithms with that of the global optimal allocation, the solution calculated by the Genetic Algorithm (GA), and the Proportional Fairness (PF) scheduler for multi-user MIMO systems as in [36]. In addition, we evaluate the computational complexity of the proposed algorithm by discussing the statistics of the major parameters that affect the complexity.



(a) H2H sum-rate versus number of HTC UEs ($A = 32,\ M = 4$).



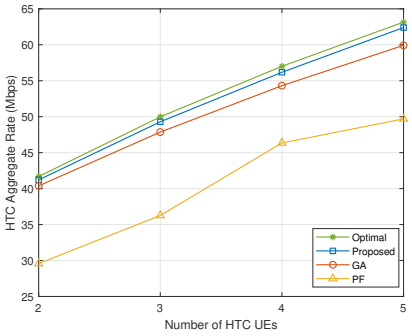
(b) HTC sum-rate versus number of MTCDs ($A = 32,\ H = 4$).



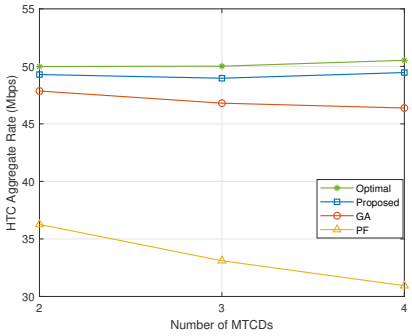
(c) HTC sum-rate versus number of antennas ($H = 6,\ M = 6$).

Fig. 4.4: Aggregate HTC achievable data rate.

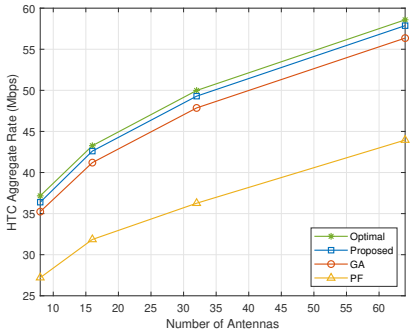
In the simulations, we uniformly distribute a set of single-antenna HTC and critical MTC users in a single LTE cell with a radius of 500 m. The users are served by a single eNB that contains a massive number of antennas which are used to simultaneously schedule more than one user on the same PRB. Without loss of generality, we use maximal ratio combining (MRC) receive beamforming vectors in the simulations. The users generate uplink transmissions with Poisson arrivals with average arrival rate uniformly picked from the sets as in Table 4.2, which summarizes the simulation parameters. As discussed in Section 4.3, we assume that the HTC UEs have



(a) Aggregate HTC data rate versus number of HTC UEs ($A = 32,\ M = 2$).

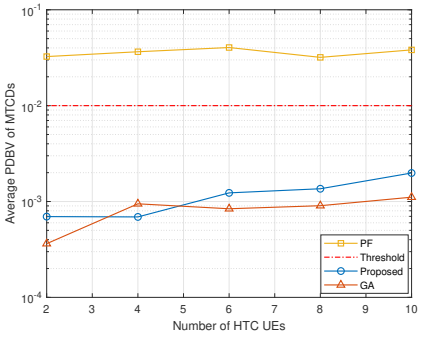


(b) Aggregate HTC data rate versus number of MTCDs ($A = 32,\ H = 3$).

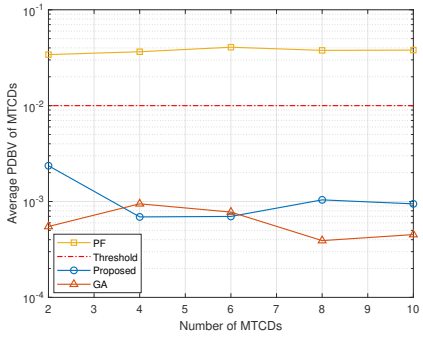


(c) Aggregate HTC data rate versus number of antennas ($H = 3,\ M = 2$).

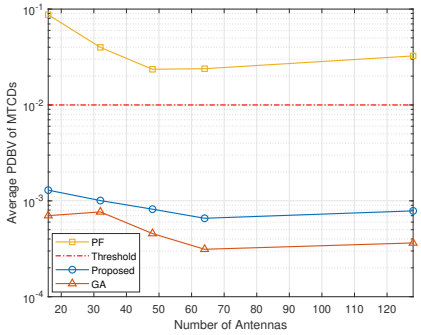
Fig. 4.5: Comparison with the global optimal solution (3 runs and 100 TTIs).



(a) Average PDBV versus number of HTC UEs ($A = 32,\ M = 4$).



(b) Average PDBV versus number of MTCDs ($A = 32,\ H = 4$).



(c) Average PDBV versus number of antennas ($H = 6,\ M = 6$).

Fig. 4.6: Average PDBV of MTCDs in the cell.

minimum average rate requirements and the critical MTCDs have minimum PDBV requirements. Therefore, the aggregate achievable data rate of the HTC UEs and the average PDBV of the MTCDs in the cell are the metrics used to evaluate the performance of the proposed resource allocation scheme. The 95% confidence interval of the estimation of the HTC aggregate rate ranges from 0.2586 Mbps to 0.7083 Mbps with an average of 0.4544 Mbps. For the estimation of the average PDBV of the MTCDs, the confidence interval varies from 1.56×10^{-4} to 2.70×10^{-3} with an average of 1.12×10^{-3}.

Figure 4.4 shows the aggregate achievable data rate of the HTC UEs in the cell using the proposed matching-based, the GA-based, and the PF schedulers. Increasing the number of HTC UEs or antennas allows the scheduler to co-schedule more HTC UEs on the same PRB. This results in an improvement in the HTC sum-rate as shown in Figs. 4.4(a) and 4.4(c). On the other hand, scheduling more MTCDs in the cell degrades the HTC sum-rate since fulfilling their QoS requirements come at the expense of the HTC data rate, as Fig. 4.4(b) reveals. In all cases, the matching-based resource allocation achieves better aggregate HTC data rate compared to the other schedulers. This is because the PF scheduler allocates the PRBs in a fair manner to all users by maximizing their data rate based on their average throughput. Nevertheless, maximizing the data rate of the MTCDs after satisfying their QoS requirements is inefficient and impacts that of the HTC as discussed in Section 4.3. On the other side, both the matching-based and the GA-based schemes maximize the data rate of the HTC UEs while satisfying the QoS requirements of all users. The GA yields a local maximum to the optimization problem but with lower objective value than the matching-based algorithm.

To show how close the solution of the matching-based algorithm is to the global optimal solution, we compare the HTC sum-rate, which is the objective function of

the optimization problem, with that of the global maximum. For this purpose, we use the BARON solver [37] to solve the optimization problem in every TTI, i.e., the problem in (4.25). BARON adopts a polyhedral branch-and-cut approach to calculate the global optimal solution of the handled optimization problem [37]. Due to the exponential computational complexity of calculating the global optimal solution of such a problem, we run the simulation on a small-size problem as demonstrated in Fig. 4.5. The figure shows the aggregate data rate of the HTC UEs in the cell versus the number of HTC UEs, the MTCDs, and the antennas. As the figure reveals, the sum-rate achieved by utilizing the matching-based algorithm is close to the global optimal rate and always better than that of the GA-based algorithm, as discussed before.

The satisfaction of the QoS requirements of the critical MTC is demonstrated in Fig. 4.6 which shows the average PDBV of the MTCDs in the cell versus the number of the HTC UEs, MTCDs, and antennas for the scheduling algorithms. As expected, both the matching-based and the GA-based algorithms satisfy the required level of QoS in all cases. This is due to the fact that any feasible solution to an optimization problem must satisfy its constraints and the constraints of the problem in (4.15) are formulated to fulfill the QoS requirements of the MTCDs. This fulfillment of the constraints could be with equality or as an inequality based on what maximizes the objective function. However, the PF scheduler targets a fair allocation on all users without considering latency requirements. Consequently, the stringent latency requirements are violated.

In addition to analyzing the computational complexity of the proposed matching-based resource allocation scheme in big-O notation, as discussed in Section 4.5.4, we calculate statistics of the major parameters affecting the complexity using simulations. For this purpose, we calculate the cumulative distribution function (CDF) of

the number of rounds, additions, deletions, and swaps performed in Algorithm 4.3. This is because these parameters mainly determine the complexity of Algorithm 4.3 that represents the major component of the complexity of the proposed scheme. Fig. 4.7 shows the CDF of the parameters after executing the matching-based scheme $28,315$ times during the simulation using different combinations of numbers of users and antennas. As the figure reveals, the maximum number of the rounds, additions, deletions, and swaps was 11, 241, 56, 68, respectively. This shows the order of those parameters and the reduced computational complexity of the proposed scheme compared to the global optimal solution that has an exponential complexity.

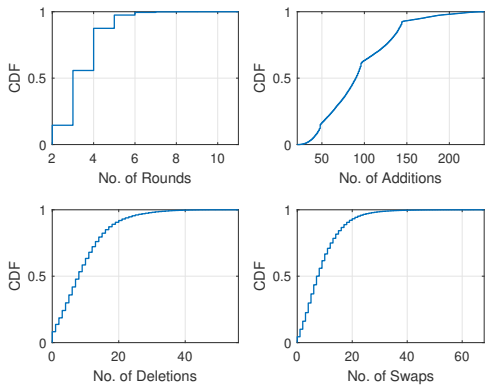


Fig. 4.7: CDF of the major parameters of the matching algorithm based on $28,315$ samples.

4.7 Conclusions

In this chapter, we utilized the effective bandwidth and effective capacity theories to formulate a cross-layer resource allocation problem for critical MTC coexistent with HTC in LTE networks with massive MIMO deployments. Then, we employed the matching theory to solve the formulated problem with much lower complexity compared to that of the global optimal solution. Therefore, the proposed matching-based

resource allocation scheme can be used in practice in LTE networks. To this end, we analyzed the computational complexity, the convergence, the stability, and the optimality of the proposed algorithms. The analysis showed that the proposed scheme converges to a local optimal allocation in a polynomial time. Extensive simulations proved the efficiency of the proposed scheme in satisfying the different types of QoS of both types of communications (HTC and critical MTC) while maximizing the system utility. The results revealed the superiority of the matching-based resource allocation compared to other algorithms of different scheduling strategies while achieving a close-to-global optimal performance. Moreover, the statistics of the major parameters that impact the computational complexity of the proposed algorithms showed the feasibility of applying the proposed scheme in practice.

References

[1] ITU-R M.2083, *IMT vision - Framework and overall objectives of the future development of IMT for 2020 and beyond*, Sept. 2015.

[2] C. Hoymann, D. Astely, M. Stattin, G. Wikstrom, J.-F. Cheng, A. Hoglund, M. Frenne, R. Blasco, J. Huschke, and F. Gunnarsson, "Lte release 14 outlook," *IEEE Communications Magazine*, vol. 54, no. 6, pp. 44–49, 2016.

[3] J. C. S. Arenas, T. Dudda, and L. Falconetti, "Ultra-low latency in next generation lte radio access," in *SCC 2017; 11th International ITG Conference on Systems, Communications and Coding; Proceedings of.* VDE, 2017, pp. 1–6.

[4] 3GPP RP-161299, *Work Item on shortened TTI and processing time for LTE*, June 2016.

[5] 3GPP RP-160667, *Work item on L2 latency reduction techniques for LTE*, March 2016.

[6] Y. Kim, H. Ji, J. Lee, Y.-H. Nam, B. L. Ng, I. Tzanidis, Y. Li, and J. Zhang, "Full dimension mimo (fd-mimo): The next evolution of mimo in lte systems," *IEEE Wireless Communications*, vol. 21, no. 2, pp. 26–33, 2014.

[7] H. Ji, Y. Kim, J. Lee, E. Onggosanusi, Y. Nam, J. Zhang, B. Lee, and B. Shim, "Overview of full-dimension mimo in lte-advanced pro," *IEEE Communications Magazine*, vol. 55, no. 2, pp. 176–184, 2017.

[8] P. Popovski, J. J. Nielsen, C. Stefanovic, E. De Carvalho, E. Strom, K. F. Trillingsgaard, A.-S. Bana, D. M. Kim, R. Kotaba, J. Park *et al.*, "Wireless access for ultra-reliable low-latency communication: Principles and building blocks," *Ieee Network*, vol. 32, no. 2, pp. 16–23, 2018.

[9] P. Popovski, Č. Stefanović, J. J. Nielsen, E. De Carvalho, M. Angjelichinoski, K. F. Trillingsgaard, and A.-S. Bana, "Wireless access in ultra-reliable low-latency communication (urllc)," *IEEE Transactions on Communications*, vol. 67, no. 8, pp. 5783–5801, August 2019.

[10] S. R. Panigrahi, N. Bjorsell, and M. Bengtsson, "Feasibility of large antenna arrays towards low latency ultra reliable communication," in *2017 IEEE International Conference on Industrial Technology (ICIT)*. IEEE, 2017, pp. 1289–1294.

[11] E. Khorov, A. Krasilov, and A. Malyshev, "Reliable low latency communications in lte networks," in *Black Sea Conference on Communications and Networking (BlackSeaCom), 2017 IEEE International*. IEEE, 2017, pp. 1–5.

[12] C. Sun, C. She, and C. Yang, "Energy-efficient resource allocation for ultra-reliable and low-latency communications," in *GLOBECOM 2017-2017 IEEE Global Communications Conference*. IEEE, 2017, pp. 1–6.

[13] C. She, C. Yang, and T. Q. Quek, "Cross-layer optimization for ultra-reliable and low-latency radio access networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 1, pp. 127–141, 2018.

[14] C. Sun, C. She, and C. Yang, "Exploiting multi-user diversity for ultra-reliable and low-latency communications," in *Globecom Workshops (GC Wkshps), 2017 IEEE*. IEEE, 2017, pp. 1–6.

[15] F. Ghavimi, Y.-W. Lu, and H.-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, 2017.

[16] A. M. Maia, D. Vieira, M. F. de Castro, and Y. Ghamri-Doudane, "A fair qos-aware dynamic lte scheduler for machine-to-machine communication," *Computer Communications*, vol. 89, pp. 75–86, 2016.

[17] M. Y. Abdelsadek, Y. Gadallah, and M. H. Ahmed, "An lte-based optimal resource allocation scheme for delay-sensitive m2m deployments coexistent with h2h users," in *2017 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, May 2017, pp. 139–144.

[18] ——, "Optimal cross-layer resource allocation for critical mtc traffic in mixed lte networks," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 6, pp. 5944–5956, June 2019.

[19] C.-S. Chang and J. A. Thomas, "Effective bandwidth in high-speed digital networks," *IEEE Journal on Selected areas in Communications*, vol. 13, no. 6, pp. 1091–1100, 1995.

[20] D. Wu and R. Negi, "Effective capacity: a wireless link model for support of quality of service," *IEEE Transactions on wireless communications*, vol. 2, no. 4, pp. 630–643, 2003.

[21] M. David, *Algorithmics Of Matching Under Preferences*. World Scientific, 2013, vol. 2.

[22] E. Björnson, M. Bengtsson, and B. Ottersten, "Optimal multiuser transmit beamforming: A difficult problem with a simple solution structure [lecture notes]," *IEEE Signal Processing Magazine*, vol. 31, no. 4, pp. 142–148, 2014.

[23] P. Popovski, "Ultra-reliable communication in 5g wireless systems," in *5G for Ubiquitous Connectivity (5GU), 2014 1st International Conference on*. IEEE, 2014, pp. 146–151.

[24] 3GPP TR 38.913, *Study on scenarios and requirements for next generation access technologies, technical specification group radio access network*, Oct. 2016.

[25] A. Mehbodniya, W. Peng, and F. Adachi, "An adaptive multiuser scheduling and chunk allocation algorithm for uplink simo sc-fdma," in *2014 IEEE International Conference on Communications (ICC)*. IEEE, 2014, pp. 2861–2866.

[26] N. Abu-Ali, A.-E. M. Taha, M. Salah, and H. Hassanein, "Uplink scheduling in lte and lte-advanced: Tutorial, survey and evaluation framework," *IEEE Communications surveys & tutorials*, vol. 16, no. 3, pp. 1239–1265, 2014.

[27] 3GPP R1-101211, *PUSCH Resource Allocation for Clustered DFT-Spread OFDM*, Feb. 2010.

[28] L. Liu and J.-F. Chamberland, "On the effective capacities of multiple-antenna gaussian channels," in *2008 IEEE International Symposium on Information Theory*. IEEE, 2008, pp. 2583–2587.

[29] J. Tang and X. Zhang, "Cross-layer-model based adaptive resource allocation for statistical qos guarantees in mobile wireless networks," *IEEE Transactions on Wireless Communications*, vol. 7, no. 6, 2008.

[30] F. Kelly, S. Zachary, and I. Ziedins, *Stochastic Networks: Theory and Applications*. Oxford University Press, 1996.

[31] J. Tang and X. Zhang, "Quality-of-service driven power and rate adaptation over wireless links," *IEEE Transactions on Wireless Communications*, vol. 6, no. 8, 2007.

[32] C. H. Papadimitriou and K. Steiglitz, *Combinatorial optimization: algorithms and complexity*. Courier Corporation, 1998.

[33] E. Bodine-Baron, C. Lee, A. Chong, B. Hassibi, and A. Wierman, "Peer effects and stability in matching markets," in *International Symposium on Algorithmic Game Theory*. Springer, 2011, pp. 117–129.

[34] D. Gale and L. S. Shapley, "College admissions and the stability of marriage," *The American Mathematical Monthly*, vol. 69, no. 1, pp. 9–15, 1962.

[35] *Further advancements for E-UTRA physical layer aspects*, 3GPP TR 36.814, 2010.

[36] L. Liu, Y.-H. Nam, and J. Zhang, "Proportional fair scheduling for multi-cell multi-user mimo systems," in *Information Sciences and Systems (CISS), 2010 44th Annual Conference on*. IEEE, 2010, pp. 1–6.

[37] M. Tawarmalani and N. V. Sahinidis, "A polyhedral branch-and-cut approach to global optimization," *Mathematical Programming*, vol. 103, no. 2, pp. 225–249, 2005.

Chapter 5

Resource Allocation in LTE with FBC and sTTIs

5.1 Abstract

Critical machine-type communications (cMTC) are targeted as a major use case in the design of the fifth generation (5G) cellular systems. In this regard, the third-generation partnership project (3GPP) has introduced several enhancements to evolve the LTE standard to meet the 5G requirements. Shortened transmission time intervals (sTTIs) are considered one of the most significant improvements proposed to satisfy the stringent latency requirements of cMTC. However, this entails several challenges to the resource allocation and scheduling process. In this chapter, we address the resource allocation and scheduling of cMTC in LTE networks. The impact on the conventional human-type communications (HTC) is considered while adopting a puncturing scheduling technique. In addition, the reliability of the cMTC is ensured by utilizing the finite blocklength coding analysis to model the transmission errors and the effective bandwidth and effective capacity concepts to guarantee the queuing

delay statistics of the cMTC packets. Moreover, we propose matching theory-based computationally efficient algorithms to solve the formulated optimal resource allocation problems with reduced complexity. The proposed methods are analyzed from a practical perspective. Extensive simulations show a close-to-optimal performance of the proposed schemes while outperforming other scheduling algorithms from the literature.

5.2 Introduction

The Internet of Things (IoT) extends the interconnectivity from humans to objects or "things". One major type of these objects requires high network availability, ultra-high reliability, very low latency, and high security. Therefore, the interconnectivity of this type of objects is known as mission-critical IoT. This enables applications such as e-health, traffic safety, emergency alarms, and industrial automation. Moreover, it unleashes many applications that could emerge if the platform that supports such type of mission-critical applications is well-established. In this regard, critical machine-type communications (cMTC), or ultra-reliable low-latency communications (URLLC), are targeted as a major use case in the design of the fifth generation (5G) cellular systems [1]. This is in addition to the massive machine-type communications and enhanced mobile broadband (eMBB). Therefore, besides the new radio (NR), the third-generation partnership project (3GPP) is evolving the long-term evolution (LTE) standard to support such type of critical communications to meet the requirements of the International Mobile Telecommunications (IMT)-2020 standard [2]. For this purpose, several enhancements in the physical (PHY) and medium access control (MAC) layers of LTE have been introduced in the latest releases of 3GPP to efficiently serve cMTC [3]. Shortened transmission time intervals (sTTIs) [4], fast uplink

access on MAC [5] and support of reduced processing times [4] are examples of these improvements.

In this regard, the support of the scheduling of cMTC at the sTTIs level is considered one of the major enhancements. This is due to the very low latency requirements of several cMTC applications that can reach the level of 0.25 ms. These stringent latency requirements are hard to be satisfied with the legacy 1 ms transmission time interval (TTI) of LTE. Therefore, with the support of sTTIs, the cMTC traffic can be scheduled on very short intervals as low as 0.143 ms based on various lengths of 2-symbol, 4-symbol, and 7-symbol transmissions [4].

Nevertheless, this support of sTTIs involves several challenges to the resource allocation and scheduling process. First, due to the short transmission intervals and small-size packets of cMTC, the finite blocklength coding (FBC) should be taken into consideration [6]. This necessitates using the analysis of the capacity of wireless channels in the finite blocklength regime as investigated in [7] instead of the conventional Shannon capacity that assumes infinite blocklength codes. Additionally, the scheduling of cMTC and conventional human-type communications (HTC) on different scales of transmission duration, i.e., sTTIs and TTIs, is a challenging task [8]. Therefore, puncturing scheduling is a technique proposed for 5G systems to efficiently support cMTC without degrading the quality of service (QoS) of HTC [9]. In puncturing scheduling, the cMTC traffic is transmitted once it arrives by pausing or overwriting the ongoing HTC transmissions based on sTTIs levels [10]. This overcomes the queuing of the cMTC packets until the next TTI, which can violate their critical latency requirements. However, the selection of the radio resources to be punctured is still a challenging task. This is due to the fact that the puncturing of the HTC resources results in a loss in their data rates. In addition, the different QoS requirements of the two types of communications should be considered while allocating the

radio resources. This necessitates the optimization of the overall resource allocation and scheduling process.

In this chapter, we address the resource allocation and scheduling of cMTC coexistent with HTC in LTE networks adopting sTTIs and considering FBC and puncturing scheduling. To the best of our knowledge, this work is the first to investigate this problem and to consider these challenges in LTE.

5.2.1 Related Work

There are several recent studies that have investigated the resource allocation problem for cMTC in cellular networks. In [11], the authors target minimizing the transmit power of the cMTC devices (cMTCDs) while considering their latency and reliability requirements in orthogonal frequency division multiple access (OFDMA) systems. This is achieved by optimizing the power allocation, bandwidth allocation, and packets dropping using FBC analysis. Similar studies in [6, 12] consider both the uplink and downlink directions while allocating the resources. In addition, the authors in [13] consider maximizing the admissible cMTC load and investigate the minimum required bandwidth. This is achieved by optimizing the resource allocation and packet re-transmission schemes. However, the aforementioned studies do not consider the impact on the HTC traffic and how the puncturing process can be optimized.

Nevertheless, the coexistence of the HTC traffic and cMTC is considered in several studies. In [14, 15], the authors address the resource allocation problem for cMTC coexistent with HTC. The objective is to maximize the data rate of the HTC users. However, they do not consider the sTTIs in LTE. The authors in [10] study punctured scheduling and the recovery mechanisms of the punctured HTC resources in addition to link adaptation and resource allocation in NR. In [16], the authors investigate the data rate loss of HTC under puncturing and categorize it into linear, convex,

and threshold models. Additionally, they study the corresponding resource allocation problems of cMTC coexistent with HTC. The authors in [17] propose a risk-sensitive based approach to minimize the impact on the HTC after puncturing their resources while satisfying reliability constraints for cMTC. However, the previously discussed studies do not consider the FBC of cMTC traffic. As discussed in [18], the queuing delay and delay-bound violation probability cannot be guaranteed if the FBC is not taken into consideration.

5.2.2 Paper Contributions and Outline

The major contributions of this chapter can be summarized as follows:

- We formulate the resource allocation of both types of traffic, i.e., HTC and cMTC, on different scales of time in LTE networks that use the puncturing scheduling technique. The aggregate data rate of the HTC users is maximized while guaranteeing them a minimum average rate. In addition, the puncturing process is optimized such that data rate loss of HTC is minimized while fulfilling the stringent reliability requirements of cMTC. In this regard, we divide the scheduling problem of cMTC into two scenarios as we discuss in Section 5.4.2. The resource allocation problem for every case is formulated and analyzed separately. Then, a unified problem is constructed for both cases. The considered reliability constraints incorporate both the transmission errors due to FBC, utilizing the analysis in [7], and the queuing-delay of the cMTC packets, using the effective bandwidth [19] and effective capacity [20] theories.

- The formulated combinatorial problems are analyzed and simplified. Then, we propose computationally-efficient algorithms to solve them. For this purpose, we utilize the matching theory [21] to formulate the resource allocation problem

as matching processes.

- The proposed matching-based scheduling schemes are analyzed from a practical perspective to be used as real-time scheduling schemes. In this regard, the convergence and stability of the proposed matchings are proved. In addition, the computational complexity of the proposed schemes is analyzed using the big-O notation. Moreover, the performance of the matching-based schedulers is evaluated and compared with other scheduling schemes using extensive simulations.

The remainder of the chapter is organized as follows. In Section 5.3, the system model is discussed. The HTC and cMTC scheduling problems are formulated and analyzed in Section 5.4. Then, the proposed matching-based schemes for HTC and cMTC are described and analyzed in Sections 5.5 and 5.6, respectively. In Section 5.7, the simulation results are presented and discussed. Finally, Section 5.8 concludes the chapter.

5.3 System Model

We consider the scheduling of the downlink transmissions of a single LTE cell with a single eNB. Assume that the cell contains a set of HTC user equipment (UEs) indexed by $\mathcal{U} = \{1, \cdots, u, \cdots, U\}$ and a set of cMTCDs indexed by $\mathcal{M} = \{1, \cdots, m, \cdots, M\}$. The set $\mathcal{K} = \{1, \cdots, k, \cdots, K\}$ represents the physical resource blocks (PRBs) to be allocated to the users such that each PRB can be used for a TTI of 1 ms by HTC UEs and for an sTTI of 0.143 ms duration by cMTCDs as shown in Fig. 5.1. The frequently used symbols are summarized in Table 5.1.

The signal-to-noise ratio (SNR) of the u^{th} HTC UE on the k^{th} PRB, $\gamma_{u,k}$, is given

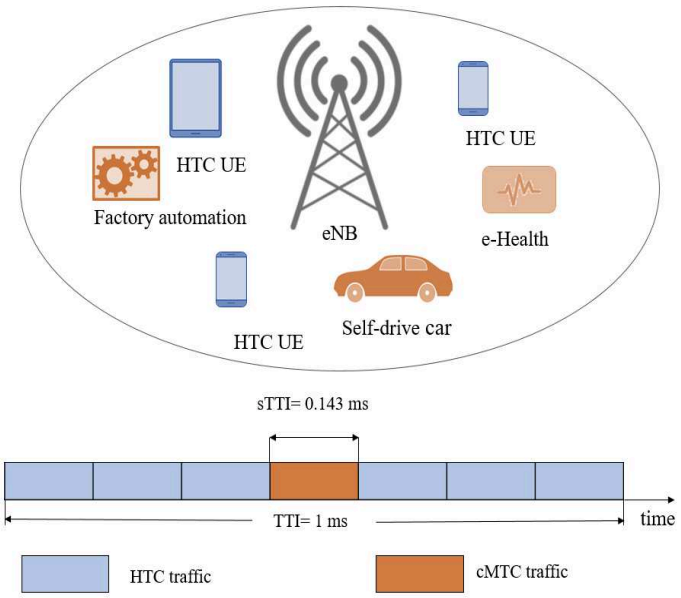


Fig. 5.1: HTC and cMTC coexistence in an LTE cell and frame structure.

by

$$\gamma_{u,k} = \frac{P_{u,k}|h_{u,k}|^2}{N_0 B}, \tag{5.1}$$

where $P_{u,k}$ is the transmit power on that PRB, $B = 180$ KHz is the bandwidth of one PRB, and N_0 is the power spectral density (PSD) of the additive white Gaussian noise (AWGN). The channel gain between the u^{th} HTC UE and the eNB on the k^{th} PRB is calculated by $h_{u,k} = \sqrt{Z_u/L_u} f_{u,k}$, where L_u is the power path loss, Z_u is the power gain due to shadowing, and $f_{u,k}$ is an independent and identically distributed (i.i.d.) complex Gaussian random variable that represents the small-scale fading.

To calculate the achievable data rate of the HTC UEs, we assume that the data rate loss on every PRB due to puncturing is linearly proportional to the number of punctured sTTIs on that PRB during a given TTI. That is, the achievable data rate

Table 5.1: Frequently Used Symbols and Notations of Chapter 5

Symbol	Description
$\mathcal{K}, \mathcal{U}, \mathcal{M}$	Sets of PRBs, HTC UEs, cMTCDs, respectively
K, U, M	Cardinalities of $\mathcal{K}, \mathcal{U}, \mathcal{M}$, respectively
$T,\ \tau$	Duration of one TTI and sTTI
$i,\ j$	Indexes of TTIs and sTTIs
B	Bandwidth of one PRB
$x,\ s$	Binary indicator variables
$R,\ R^{min},\ \bar{R}^{min}$	Data rate, minimum rate, minimum average rate
γ	signal-to-noise ratio
D_m^{max}	Delay bound of mth cMTCD
ϵ_m^{max}	Maximum allowed PTE of mth cMTCD
ε_m^{max}	Maximum allowed PDBV of mth cMTCD
θ_m	QoS exponent of mth cMTCD
Ω_m	Effective bandwidth of mth cMTCD
Ξ_m	Effective capacity of mth cMTCD
μ^{UK}	Assignment operation of the HTC matching
μ^{MK}	Assignment operation of the cMTC matching

of the u^{th} HTC UE on the k^{th} PRB during a certain TTI is calculated by

$$R_{u,k} = B(1 - \frac{\varrho_k \tau}{T}) \log_2(1 + \gamma_{u,k}), \tag{5.2}$$

where T and τ are the duration of one TTI and sTTI, respectively, and ϱ_k is the number of punctured sTTIs on the k^{th} PRB during that TTI.

On the other hand, we take into consideration the FBC to calculate the achievable data rate of the cMTCDs. This is due to the short transmission interval of the small-size packets of the cMTC traffic. According to [6, 7], the achievable data rate of the

mth cMTCD in the finite blocklength regime can be accurately approximated by

$$R_m \approx \frac{B|\mathcal{K}_m|}{\ln 2}\left[\ln(1+\gamma_m) - \sqrt{\frac{V_m}{\tau B|\mathcal{K}_m|}}Q^{-1}(\epsilon_m)\right], \tag{5.3}$$

$$V_m = 1 - \frac{1}{(1+\gamma_m)^2}, \tag{5.4}$$

$$\gamma_m = \frac{P_m|h_m|^2}{N_0 B|\mathcal{K}_m|}, \tag{5.5}$$

where $\mathcal{K}_m$ is the set of PRBs assigned to the mth cMTCD and $|\mathcal{K}_m|$ is its cardinality; Q^{-1} is the inverse of the Q-function; and ϵ_m is the probability of transmission error (PTE).

5.4 Problem Formulation and Analysis

We now formulate the resource allocation problem for the HTC traffic at every TTI and that of cMTC traffic at every sTTI. In addition, we describe the strategies and techniques used to simplify the resulting combinatorial problems such that the optimal solution can be calculated with lower complexity.

5.4.1 Scheduling of the HTC Traffic

The HTC traffic is scheduled every TTI assuming that $\varrho_k = 0$ in the next TTI. Therefore, the resource allocation problem at every TTI can be formulated as follows:

$$\max_{\mathcal{K}_u} \quad \sum_{u=1}^{U} R_u \tag{5.6}$$

$$\text{s.t.} \quad \mathbb{E}\{R_u\} \geq \bar{R}_u^{min}, \ \forall u \in \mathcal{U} \tag{5.6a}$$

$$\mathcal{K}_u \cap \mathcal{K}_{u'} = \varnothing, \ \forall u \neq u', u, u' \in \mathcal{U}, \tag{5.6b}$$

where $\mathcal{K}_u$ is the set of PRBs allocated to the u^{th} HTC UE, R_u is the data rate achieved over it, and $\mathbb{E}\{R_u\}$ is the average data rate. The constraint in (5.6a) is used to guarantee an average data rate, $\bar{R}_u^{min}$, for every HTC UE. Constraint (5.6b) is expressed to make sure that every PRB is allocated to only one user. Therefore, in the scheduling process of the HTC traffic, we aim to maximize the system utility in terms of the aggregate data rate of the HTC UEs while guaranteeing a certain average rate for every HTC UE to satisfy the heterogeneous QoS requirements of the users.

Accordingly, the average achievable data rate of every HTC UE in every TTI should be calculated to ensure that it is at least equal to the required level up to the current TTI. For this purpose, we use Lemma 5.4.1 to express the long-term requirement as an instantaneous constraint during TTIs.

Lemma 5.4.1. *To fulfill the average rate constraint of the HTC UEs as expressed in (5.6a), the instantaneous rate in the ith TTI should satisfy:*

$$R_u[i] \geq R_u^{min}[i], \ \forall u \in \mathcal{U}, \tag{5.7}$$

$$R_u^{min}[i] = i\bar{R}_u^{min} - (i-1)R_u^{avg}[i-1], \tag{5.8}$$

$$R_u^{avg}[i] = \begin{cases} \frac{R_u[i]+(i-1)R_u^{avg}[i-1]}{i}, & i \geq 2 \\ R_u[i], & i = 1 \end{cases}, \tag{5.9}$$

where $R_u^{min}[i]$ represents the required minimum instantaneous data rate in the ith TTI for the u^{th} HTC UE.

Proof. To ensure that the average data rate constraint is satisfied up to the current TTI, we use a moving-average expression that is calculated every TTI. For this purpose, the cumulative moving average (CMA) as expressed in (5.9) can be used to relate the current instantaneous rate to the previous allocated rates. This represents a valid estimation of the average value of the data rate due to the ergodicity of the

random process composed by the sequence $\{R_u[i] : \ i = 1, 2, 3, \cdots\}$. Therefore, the average data rate constraint in (5.6a) can be written as

$$R_u^{avg}[i] \geq \bar{R}_u^{min}, \ \forall u \in \mathcal{H}. \tag{5.10}$$

Thus, by solving this inequality for $R_u[i]$ using (5.9), it can be formulated as

$$R_u[i] \geq i\bar{R}_u^{min} - (i-1)R_u^{avg}[i-1], \ \forall u \in \mathcal{U}. \tag{5.11}$$

This can be written in the form as in (5.7) and (5.8). □

Therefore, using a binary variable $x_{u,k}$ to indicate whether PRB k is assigned to the u^{th} HTC UE, the resource allocation problem in (5.6) can be formulated as:

$$\max_{\mathbf{X}} \ \sum_{u=1}^{U}\sum_{k=1}^{K} x_{u,k}R_{u,k}[i] \tag{5.12}$$

$$\text{s.t.} \ \sum_{k=1}^{K} x_{u,k}R_{u,k}[i] \geq R_u^{min}[i], \ \forall u \in \mathcal{U} \tag{5.12a}$$

$$\sum_{u=1}^{U} x_{u,k} \leq 1, \ \forall k \in \mathcal{K} \tag{5.12b}$$

$$x_{u,k} \in \{0, 1\}, \ \forall k \in \mathcal{K}, \ u \in \mathcal{U}, \tag{5.12c}$$

where $\mathbf{X}$ is a $U \times K$ indicator matrix. Constraints (5.12a) and (5.12b) are equivalent to (5.6a) and (5.6b), respectively. Constraint (5.12c) is used to restrict the decision variable $x_{u,k}$ to binary values. This optimization problem can be modeled as a binary linear program (BLP) which can be solved optimally using algorithms such as Branch and Bound. However, the computational complexity is too high to be executed in real-time every TTI. Therefore, we propose polynomial-time algorithms to solve this problem as discussed in Section 5.5.

5.4.2 Scheduling of the cMTC Traffic

Adopting a puncturing scheduling technique, the cMTC traffic can be transmitted by pausing the ongoing HTC transmissions for an sTTI period to avoid queuing the cMTC packets until the next TTI. Therefore, the cMTC packets can be totally transmitted in the first sTTI that follows their arrival or can be queued for a few sTTIs as long as the queueing delay is guaranteed to be within a certain limit. We first consider the two scenarios separately and then formulate a unified problem for the two cases.

5.4.2.1 Case 1: Immediate Transmission of the cMTC Traffic

In this scenario, once a cMTC packet arrives, it is transmitted immediately in the next sTTI by puncturing an ongoing HTC transmission. However, due to the FBC, the reliability requirements of cMTC, in terms of the PTE, should be taken into consideration. Therefore, in the resource allocation process, the bandwidth of every cMTCD is assigned such that the PTE of every device satisfies

$$\epsilon_m \leq \epsilon_m^{max}, \tag{5.13}$$

where ϵ_m^{max} is the maximum allowed PTE of the mth device. The data rate of the mth device in (5.3) can be expressed as

$$\frac{\beta_m}{\tau} \approx \frac{B|\mathcal{K}_m|}{\ln 2}\left[\ln(1+\gamma_m) - \sqrt{\frac{V_m}{\tau B|\mathcal{K}_m|}}Q^{-1}(\epsilon_m)\right], \tag{5.14}$$

where β_m is its packet size. Accordingly, we can rewrite (5.13) in the following form

$$Q\left(\frac{\ln(1+\gamma_m) - \frac{\beta_m \ln 2}{\tau B|\mathcal{K}_m|}}{\sqrt{\frac{V_m}{\tau B|\mathcal{K}_m|}}}\right) \leq \epsilon_m^{max}. \tag{5.15}$$

In addition, the selection of the PRBs to be punctured should be optimized such that the data rate loss of HTC is minimized and their QoS requirements are not violated. Therefore, the resource allocation problem at every sTTI can be formulated as

$$\min_{\mathbf{S}} \quad \sum_{k=1}^{K}\sum_{m=1}^{M} s_{m,k}\sum_{u=1}^{U} x_{u,k}\frac{\tau B}{T}\log_2(1+\gamma_{u,k}) \tag{5.16}$$

$$\text{s.t.} \quad Q\left(\frac{\ln(1+\frac{P_m|h_m|^2}{N_0 B\sum_{k=1}^{K} s_{m,k}}) - \frac{\beta_m \ln 2}{\tau B\sum_{k=1}^{K} s_{m,k}}}{\sqrt{\frac{V_m}{\tau B\sum_{k=1}^{K} s_{m,k}}}}\right) \leq \epsilon_m^{max},\ \forall m \in \mathcal{M} \tag{5.16a}$$

$$\sum_{k=1}^{K} x_{u,k}\left(1-\frac{\tau}{T}\left(\varrho_k - \sum_{m=1}^{M} s_{m,k}\right)\right) B\log_2(1+\gamma_{u,k}) \geq R_u^{min}[i],\ \forall u \in \mathcal{U} \tag{5.16b}$$

$$\sum_{m=1}^{M} s_{m,k} \leq 1,\ \forall k \in \mathcal{K} \tag{5.16c}$$

$$s_{m,k} \in \{0,1\},\ \forall k \in \mathcal{K},\ m \in \mathcal{M}, \tag{5.16d}$$

where $\mathbf{S}$ is an $M \times K$ binary indicator matrix. Constraints (5.16a) and (5.16b) are used to fulfill the reliability requirements of cMTC and the data rate requirements of HTC, respectively, where ϱ_k represents the number of puncturing operations on the k^{th} PRB during the current TTI up to the last sTTI. Constraint (5.16c) is used to ensure that every PRB is punctured by only one cMTCD, at most. The binary value of the indicator variable, $s_{m,k}$, is restricted in constraint (5.16d).

5.4.2.2 Case 2: Transmission of cMTC Traffic with Queueing

In this case, the cMTC packets are queued on the sTTIs level such that the packet loss due to queueing is ensured to be under a certain threshold as depicted in Fig. 5.2. To provide guarantees for the queueing delay of the cMTCDs, the effective bandwidth [19] and effective capacity [20] theories provide a powerful approach for statistical QoS guarantees for time-varying buffer dynamics. In this regard, the statistical guarantees

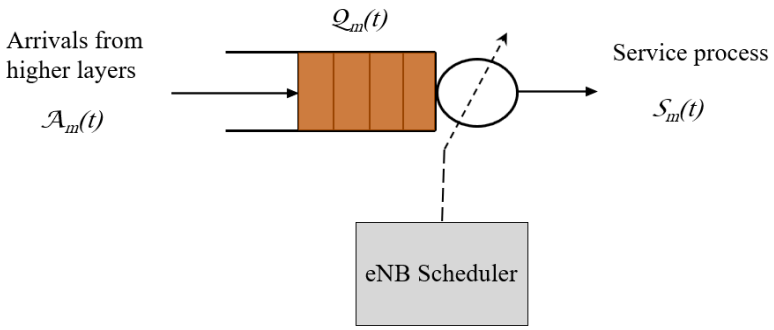


Fig. 5.2: Queueing model at the eNB.

are expressed in terms of the probability of delay-bound violation (PDBV) in the following form

$$\Pr[D_m(t) \geq D_m^{max}] \leq \varepsilon_m^{max}, \tag{5.17}$$

where $D_m(t)$ and D_m^{max} are the queueing delay at time t and the delay-bound of the mth device, respectively, and ε_m^{max} represents the maximum allowed packet loss due to queuing, assuming that packets that exceed their deadlines are dropped. Therefore, this is counted as another source of losses along with that associated with the transmission errors that are restricted by ϵ_m^{max}.

According to the large deviations and effective bandwidth theories [19], the queue length process, $\mathcal{Q}_m(t)$, converges in distribution to a random variable $\mathcal{Q}_m(\infty)$ such that

$$-\lim_{Q_m^{th}\to\infty} \frac{\ln\left(\Pr[\mathcal{Q}_m(\infty) < Q_m^{th}]\right)}{Q_m^{th}} = \theta_m, \tag{5.18}$$

where Q_m^{th} is the queue-length bound and $\theta_m > 0$ is known as the QoS exponent since

it determines the queue-length decaying rate such that

$$\Pr[D_m(t) \geq D_m^{max}] \approx e^{-\theta_m \delta_m D_m^{max}}, \tag{5.19}$$

where δ_m is a parameter that is determined by the arrival and service processes. For example, δ_m for a Poisson process can be calculated as [22]

$$\delta_m = \lambda_m \left(\frac{e^{\theta_m} - 1}{\theta_m} \right), \tag{5.20}$$

where λ_m is the average arrival rate.

In this respect, the effective bandwidth of an arrival process, $\mathcal{A}_m(t)$, is defined as the minimum constant service rate that can serve the process with a QoS exponent θ_m and can be calculated by [19]

$$\Omega_m(\theta_m) = \lim_{t \to \infty} \frac{1}{t\theta_m} \ln \mathbb{E}\{e^{\theta_m \mathcal{A}_m(t)}\}. \tag{5.21}$$

A corresponding concept discussed in [20] known as the effective capacity is defined as the maximum constant arrival rate that can be served by a service process $\mathcal{S}_m(t)$ with a guaranteed QoS exponent θ_m, and is calculated by

$$\Xi_m(\theta_m) = -\lim_{t \to \infty} \frac{1}{t\theta_m} \ln \mathbb{E}\{e^{-\theta_m \mathcal{S}_m(t)}\}. \tag{5.22}$$

In this context, the QoS exponent, θ_m, represents the required queueing delay behavior. It can be calculated from D_m^{max} and ε_m^{max} using (5.17) and (5.19) as

$$\theta_m = \frac{-\ln \varepsilon_m^{max}}{\delta_m D_m^{max}}. \tag{5.23}$$

Consequently, to guarantee a certain PDBV for a device as in (5.17), the effective

capacity and effective bandwidth should satisfy

$$\Xi_m(\theta_m) \geq \Omega_m(\theta_m). \tag{5.24}$$

In the considered case of the buffers of the devices in the eNB, the service process can be defined as

$$\mathcal{S}_m[t] = \sum_{j=1}^{t} R_m[j]\tau, \tag{5.25}$$

where j is the sTTI index such that $\{R_m[j]\tau : j = 1, 2, 3, \cdots\}$ is a discrete-time stationary and ergodic random process. Therefore, given that the sequence $\{R_m[j]\tau : j = 1, 2, 3, \cdots\}$ is uncorrelated, the effective capacity of $\mathcal{S}_m[t]$ can be derived as

$$\begin{aligned}
\Xi_m(\theta_m) &= -\lim_{t\to\infty} \frac{1}{t\theta_m} \ln \mathbb{E}\{e^{-\theta_m \sum_{j=1}^{t} R_u[j]\tau}\} && (5.26)\\
&= -\lim_{t\to\infty} \frac{1}{t\theta_m} \ln \mathbb{E}\left\{\prod_{j=1}^{t} e^{-\theta_m R_m[j]\tau}\right\} && (5.27)\\
&= -\lim_{t\to\infty} \frac{1}{t\theta_m} \ln \left(\mathbb{E}\{e^{-\theta_m R_m[j]\tau}\}\right)^t && (5.28)\\
&= \frac{-1}{\theta_m} \ln \mathbb{E}\{e^{-\theta_m R_m[j]\tau}\}. && (5.29)
\end{aligned}$$

Therefore, the required QoS constraint in (5.24) can be expressed as

$$\frac{-1}{\theta_m} \ln \mathbb{E}\{e^{-\theta_m R_m[j]\tau}\} \geq \Omega_m. \tag{5.30}$$

Moreover, this constraint can be enforced every sTTI by applying Theorem 5.4.2 as follows.

Theorem 5.4.2. *To fulfill the effective capacity constraint in (5.30), the instanta-*

neous data rate in the jth sTTI should satisfy:

$$R_m[j] \geq R_m^{min}[j], \ \forall m \in \mathcal{M}, \tag{5.31}$$

$$R_m^{min}[j] = \frac{-1}{\theta_m \tau} \ln \left(j e^{-\theta_m \Omega_m} - (j-1) \Upsilon_m^{avg}[j-1] \right), \tag{5.32}$$

$$\Upsilon_m^{avg}[j] = \begin{cases} \frac{\Upsilon_m[j] + (j-1)\Upsilon_m^{avg}[j-1]}{i}, & j \geq 2 \\ \Upsilon_m[j], & j = 1 \end{cases}, \tag{5.33}$$

$$\Upsilon_m[j] = e^{-\theta_m R_m[j]\tau}. \tag{5.34}$$

Proof. To make sure that the constraint in (5.30) is satisfied until the jth sTTI we use the cumulative moving average to estimate $\Upsilon_m[j]$ defined in (5.34). In this regard, we define $\Upsilon_m^{avg}[j]$ as the CMA of $\Upsilon_m[j]$ at the jth sTTI and is calculated as given in (5.33). This represents a valid estimation of $\mathbb{E}\{e^{-\theta_m R_m[j]\tau}\}$ since the random process $\{R_m[j]\tau : \ j = 1, 2, 3, \cdots\}$ is stationary and ergodic. Therefore, the constraint in (5.30) can be rewritten as

$$\Upsilon_m^{avg}[j] \leq e^{-\theta_m \Omega_m}. \tag{5.35}$$

Using (5.33), (5.35) can be reordered as

$$\Upsilon_m[j] = e^{-\theta_m R_m[j]\tau} \leq j e^{-\theta_m \Omega_m} - (j-1)\Upsilon_m^{avg}[j-1]. \tag{5.36}$$

Therefore,

$$R_m[j] \geq \frac{-1}{\theta_m \tau} \ln \left(j e^{-\theta_m \Omega_m} - (j-1) \Upsilon_m^{avg}[j-1] \right). \tag{5.37}$$

This can be formulated as in (5.31) and (5.32). □

In addition to the PDBV constraint, cMTC also requires guarantees for the PTE as in (5.13). Therefore, from (5.31) and (5.3), we can combine the PDBV and PTE requirements in the constraint in (5.38).

$$\frac{B}{\ln 2}\sum_{k=1}^{K} s_{m,k}\left[\ln\left(+\frac{P_m|h_m|^2}{N_0 B\sum_{k=1}^{K} s_{m,k}}\right)-\sqrt{\frac{V_m}{\tau B\sum_{k=1}^{K} s_{m,k}}}Q^{-1}(\epsilon_m^{max})\right]\geq R_m^{min}[j],\ \forall m\in\mathcal{M} \tag{5.38}$$

Accordingly, the resource allocation problem for cMTC at every sTTI is formulated such that the rate loss of HTC is minimized while satisfying the QoS requirements as follows

$$\max_{\mathbf{S}}\ \sum_{k=1}^{K}\sum_{m=1}^{M} s_{m,k}\sum_{u=1}^{U} x_{u,k}\frac{\tau B}{T}\log_2(1+\gamma_{u,k}) \tag{5.39}$$

$$\text{s.t.}\ \ (5.38),\ (5.16b),\ (5.16c),\ (5.16d).$$

5.4.2.3 Unified Problem for Cases 1 and 2

The problems of cMTC scheduling in (5.16) and (5.39) can be reformulated and simplified in a unified problem as follows. The difference between the problems is the reliability constraints, i.e., (5.38) and (5.16a). Therefore, we combine these two constraints in a single general one. This can be achieved given that the two constraints are functions of $\sum_{k=1}^{K} s_{m,k}$, i.e., the number of assigned PRBs for every device. Therefore, an equivalent constraint can be expressed as

$$\sum_{k=1}^{K} s_{m,k}\geq K_m^{min}, \tag{5.40}$$

where K_m^{min} is the minimum number of PRBs that is required by the mth device to satisfy its PTE requirement in Case 1, or PTE and PDBV requirements in Case 2.

This can be calculated numerically from (5.38) and (5.16a) with low complexity due to the limited number of PRBs as will be analyzed in Section 5.6.

Consequently, the resulting optimization problem can be reformulated in a BLP form as follows. First, we formulate the constraint in (5.16b) to be in a linear form by rearranging it as

$$\sum_{k=1}^{K} x_{u,k} \frac{B\tau}{T} \log_2(1+\gamma_{u,k}) \sum_{m=1}^{M} s_{m,k} \;\leq \sum_{k=1}^{K} x_{u,k} \left(1 - \frac{\varrho_k \tau}{T}\right) B \log_2(1+\gamma_{u,k}) - R_u^{min}[i], \forall u \in \mathcal{U}. \tag{5.41}$$

Then, we can simplify (5.41) to be in the following form

$$\sum_{k=1}^{K} \sum_{m=1}^{M} R_{u,k}^{P} s_{m,k} \leq R_u^{allow}, \tag{5.42}$$

where,

$$R_{u,k}^{P} = x_{u,k} \frac{B\tau}{T} \log_2(1+\gamma_{u,k}), \tag{5.43}$$

$$R_u^{allow} = \sum_{k=1}^{K} x_{u,k} (1 - \frac{\varrho_k \tau}{T}) B \log_2(1+\gamma_{u,k}) - R_u^{min}[i]. \tag{5.44}$$

Accordingly, the unified cMTC scheduling problem can be expressed as

$$\max_{\mathbf{S}} \quad \sum_{k=1}^{K}\sum_{m=1}^{M} s_{m,k} \sum_{u=1}^{U} x_{u,k} \frac{\tau B}{T} \log_2(1+\gamma_{u,k}) \tag{5.45}$$

$$\text{s.t.} \quad \sum_{k=1}^{K} s_{m,k} \geq K_m^{min},\ \forall m \in \mathcal{M} \tag{5.45a}$$

$$\sum_{k=1}^{K}\sum_{m=1}^{M} R_{u,k}^{P} s_{m,k} \leq R_u^{allow},\ \forall u \in \mathcal{U} \tag{5.45b}$$

$$\sum_{m=1}^{M} s_{m,k} \leq 1,\ \forall k \in \mathcal{K} \tag{5.45c}$$

$$s_{m,k} \in \{0,1\},\ \forall k \in \mathcal{K},\ m \in \mathcal{M}. \tag{5.45d}$$

This optimization problem can be written in the following BLP form

$$\min_{\tilde{\mathbf{s}}} \quad \mathbf{c}^T \tilde{\mathbf{s}} \tag{5.46}$$

$$\text{s.t.} \quad \mathbf{A}\tilde{\mathbf{s}} \leq \mathbf{b} \tag{5.46a}$$

$$\tilde{s} \in \{0,1\}, \tag{5.46b}$$

where $\tilde{\mathbf{s}},\ \mathbf{c},\ \mathbf{A}$, and $\mathbf{b}$ are constructed as follows.

The decision variable, $\tilde{\mathbf{s}}$, is related to $\mathbf{S}$ as

$$\tilde{\mathbf{s}} = [\mathbf{s}_1\ \cdots\ \mathbf{s}_m\ \cdots\ \mathbf{s}_M]^T, \tag{5.47}$$

where $\mathbf{s}_m$ is the mth row of the matrix $\mathbf{S}$. The cost vector, $\mathbf{c} = [\mathbf{R}^l\ \cdots\ \mathbf{R}^l]^T$, is composed of M-duplicates of $\mathbf{R}^l$ that is calculated as

$$\mathbf{R}^l = [R_1^l\ \cdots\ R_k^l\ \cdots\ R_K^l], \tag{5.48}$$

$$R_k^l = \sum_{u=1}^{U} x_{u,k} \frac{\tau B}{T} \log_2(1+\gamma_{u,k}). \tag{5.49}$$

The constraints matrix and vector can be derived as follows. The right-hand-side of the inequality, $\mathbf{b}$, includes the three constraints in (5.45a), (5.45b), and (5.45c) and is calculated by

$$\mathbf{b} = [-K_1^{min} \ \cdots \ -K_M^{min} \ R_1^{allow} \ \cdots \ R_U^{allow} \ \mathbf{1}_K^T]^T, \tag{5.50}$$

where $\mathbf{1}_K$ is the K-length ones vector. In the same manner, the constraints matrix, $\mathbf{A}$, is composed of three parts as

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}^a \\ \mathbf{A}^b \\ \mathbf{A}^c \end{pmatrix}, \tag{5.51}$$

where $\mathbf{A}^a$, $\mathbf{A}^b$, and $\mathbf{A}^c$ represent the left-hand-side of the constraints in (5.45a), (5.45b), and (5.45c), respectively. The first component, $\mathbf{A}^a$, is an $M \times MK$ matrix that is composed as

$$\mathbf{A}^a = \begin{pmatrix} -\mathbf{1}_K^T & \mathbf{0}_K^T & \cdots & \mathbf{0}_K^T \\ \mathbf{0}_K^T & -\mathbf{1}_K^T & \cdots & \mathbf{0}_K^T \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_K^T & \mathbf{0}_K^T & \cdots & -\mathbf{1}_K^T \end{pmatrix}, \tag{5.52}$$

where $\mathbf{0}_K$ is the K-length zeroes vector. The second part, $\mathbf{A}^b$, is a $U \times MK$ matrix

that is constructed as

$$\mathbf{A}^b = \begin{pmatrix} \mathbf{R}_1^P & \mathbf{R}_1^P & \cdots & \mathbf{R}_1^P \\ \mathbf{R}_2^P & \mathbf{R}_2^P & \cdots & \mathbf{R}_2^P \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{R}_U^P & \mathbf{R}_U^P & \cdots & \mathbf{R}_U^P \end{pmatrix}, \tag{5.53}$$

where $\mathbf{R}_u^P = [R_{u,1}^P \ \cdots \ R_{u,K}^P]$. Finally, the last part $\mathbf{A}^c$ is a $K \times MK$ matrix that is composed of M duplicates of an identity matrix of size K, $\mathbf{I}_K$, as follows

$$\mathbf{A}^c = (\mathbf{I}_K \ \mathbf{I}_K \ \cdots \ \mathbf{I}_K). \tag{5.54}$$

Therefore, the optimal solution of the resulting BLP problem can be calculated with much lower complexity compared to the BNLP problems in (5.16) and (5.39).

5.5 Matching-Based Scheduling for HTC Traffic

In this section, we propose a polynomial time, computationally-efficient scheduling algorithm for the HTC traffic utilizing the matching theory. Then, we analyze the proposed algorithm from a practical point of view.

5.5.1 Matching Setup and Algorithm

The resource allocation problem in (5.12) can be formulated as a two sided-matching process assuming that $\mathcal{U}$ and $\mathcal{K}$ are two disjoint sets of agents. By ordering the agents of the opposite set, every agent $u \in \mathcal{U}$, or $k \in \mathcal{K}$, composes a preference list, $\mathcal{P}^U(u)$, or $\mathcal{P}^K(k)$, of the agents that it is willing to be matched to. This means that a PRB k is more preferred than k' by user u, expressed as $k \succ_u k'$, if it precedes k' in the

preference list $\mathcal{P}^U(u)$. Also, $u \succ_k u'$ if u precedes u' in $\mathcal{P}^K(k)$. The preferences over agents are transitive which means that if $k \succ_u k'$ and $k' \succ_u k''$, then $k \succ_u k''$.

To match the agents to each other, we design an assignment operation, μ^{UK}, such that the resource allocation problem in (5.12) is solved. For this purpose, we use the following definition of the matching of this process.

Definition 5.5.1. *To solve the resource allocation problem in (5.12), a matching μ^{UK} can be defined as a mapping from the set $\mathcal{U} \cup \mathcal{K}$ into the set $\mathcal{U} \cup \mathcal{K}$ such that for any $u \in \mathcal{U}$ and $k \in \mathcal{K}$*

(i) $\mu^{UK}(u) \subseteq \mathcal{K}$,

(ii) $\mu^{UK}(k) \in \mathcal{U}$,

(iii) $|\mu^{UK}(k)| \leq 1$,

(iv) $|\mu^{UK}(u)| \geq q_u^{min}$,

(v) $k \in \mu^{UK}(u)$ *if and only if* $\mu^{UK}(k) = u$.

Condition (i) is used to make sure that every $u \in \mathcal{U}$ can be matched to several PRBs. However, conditions (ii) and (iii) are expressed to indicate that every $k \in \mathcal{K}$ can be matched to only one u. For the minimum rate constraint of every HTC user, condition (iv) is used, where q_u^{min} is the cardinality of the set of matched PRBs that satisfy that constraint. Finally, condition (v) ensures that a PRB k is in the match list of a certain user u if and only if that user is in the match list of that PRB k.

This assignment is modeled as a one-to-many two-sided matching process. It is similar to the hospitals-residents problem with lower and higher quota bounds, as in [23]. However, the lower quota bound for every user is determined based on the set of the matched PRBs, i.e., the data rate achieved over the matched PRBs should satisfy the minimum rate constraint. Moreover, there is no higher quota bound. This

Algorithm 5.1 Matching-Based HTC Scheduling Algorithm

Step 1: Initial setup

Calculate $R_u^{min}[i]$ for all $u \in \mathcal{U}$ using (5.7).

Construct $\mathcal{P}^K(k)$ of all $k \in \mathcal{K}$ over $u \in \mathcal{U}$ and $\mathcal{P}^U(u)$ of all $u \in \mathcal{U}$ over $k \in \mathcal{K}$, based on $R_{u,k}$.

Step 2: Initial matching

Match every $k \in \mathcal{K}$ to its most preferred u in $\mathcal{P}^K(k)$.

Construct the set $\mathcal{U}^{unsat}$ of unsatisfied users, that have their $R_u^{min}[i]$ unsatisfied in the initial matching.

Based on the achievable rate on them in the initial matching, order the PRBs in an ascending manner in a set $\mathcal{P}^{comU}$, which is a temporary common preference list for all $u \in \mathcal{U}^{unsat}$ over all $k \in \mathcal{K}$.

Step 3: Matching process

Set the status of every $u \in \mathcal{U}^{unsat}$ to 1, where a status of 1 indicates that the user is still unsatisfied and otherwise is status 0.

while any user's status is 1 **do**

Feasibility test

  if $\mathcal{P}^{comU}$ is empty **then**

    Problem is infeasible. **Stop**

  end if

  $k^{propose} \leftarrow$ first k in $\mathcal{P}^{comU}$.

  $u^{prefered} \leftarrow$ most preferred u in $\mathcal{U}^{unsat}$.

  Remove $k^{propose}$ from $\mathcal{P}^{comU}$.

  if the user initially matched to $k^{propose}$ approves the unmatch based on its current rate **then**

    Match $k^{propose}$ to $u^{prefered}$.

    Update $\mu^{UK}(k^{propose})$, $\mu^{UK}(u^{prefered})$, and that of the old match of $k^{propose}$.

    if $u^{prefered}$ is now satisfied **then**

      Set its status to 0 and remove it from $\mathcal{U}^{unsat}$.

    end if

  end if

end while

makes the matching process more challenging. Consequently, we propose Algorithm 5.1 as a matching technique for this problem.

5.5.2 Matching Analysis

To analyze the proposed matching algorithm from a practical perspective, we prove its convergence and stability. Then, its computational complexity is analyzed.

The convergence of the matching in Algorithm 5.1 can be proved using the following lemma.

Lemma 5.5.2. *The proposed matching scheme in Algorithm 5.1 converges to a final matching after a finite number of iterations.*

Proof. In Algorithm 5.1, the PRBs in $\mathcal{P}^{comU}$ propose to the users $u \in \mathcal{U}^{unsat}$ that are not satisfied from the initial matching step until one of the sets, $\mathcal{U}^{unsat}$ or $\mathcal{P}^{comU}$, becomes empty. Every proposing PRB, $k^{propose}$, is removed from the list $\mathcal{P}^{comU}$ after the proposal process. This means that the proposing process should stop after a limited number of iterations since the number of PRBs in the list $\mathcal{P}^{comU}$ is finite. Consequently, the matching process μ^{UK} as described in Algorithm 5.1 converges to a final matching after a finite number of iterations. □

As discussed in [21], a two-sided matching is said to be stable if it admits no blocking pair. Therefore, we first discuss the conditions of a blocking pair in the following definition.

Definition 5.5.3. *For the considered matching process, μ^{UK}, as defined in Definition 5.5.1, a pair of agents (u', k') is called a blocking pair if:*

1. $\mu^{UK}(k') \neq u',\ k' \notin \mu^{UK}(u')$,
2. $u' \succ_{k'} \mu^{UK}(k')$,
3. $k' \succ_{u'} k'',\ k'' \in \mu^{UK}(u')$, *and*
4. *the lower quota bounds of $\mu^{UK}(k')$ and u' would still be satisfied if u' is matched to k'.*

Therefore, the stability of the matching process described in Algorithm 5.1 can be proved based on this definition of a blocking pair as follows.

Theorem 5.5.4. *Based on the definition of a blocking pair as in Definition 5.5.3, the matching scheme in Algorithm 5.1 admits no blocking pair and is stable accordingly.*

Proof. In the initial matching step in Algorithm 5.1, the PRBs are matched to their most preferred users based on their preference lists. Then, these PRBs are ordered in the set $\mathcal{P}^{comU}$ based on the data rate achieved over them by the initial match. After that, they propose in order to the most preferred users in the set $\mathcal{U}^{unsat}$. This means that the PRBs are matched to their most preferred user in Step 2 or Step 3. Also, the ordering of the PRBs in the set $\mathcal{P}^{comU}$ makes the PRBs that are less preferred by their initial match are considered first in Step 3. Therefore, let us assume that (u', k') is a pair that satisfies $\mu^{UK}(k') \neq u',\ k' \notin \mu^{UK}(u'), k' \succ_{u'} k'',\ k'' \in \mu^{UK}(u')$, and the quota bounds of $\mu^{UK}(k')$ and u' would still be fulfilled if u' is matched to k'. If the current match of k' is due to the initial matching phase, then $u' \not\succ_{k'} \mu^{UK}(k')$ because k' is matched to its most preferred user in the initial phase. The other possibility is that the current match of k' is due to the second phase of matching in Step 3. If that is the case, then $u' \not\succ_{k'} \mu^{UK}(k')$ as well since k' proposes to its most preferred $u \in \mathcal{U}^{unsat}$ in Step 3. Therefore, according to Definition 5.5.3, (u', k') cannot be considered as a blocking pair. Consequently, Algorithm 5.1 admits no blocking pair and is stable accordingly. □

For the analysis of the computational complexity of the proposed algorithm, we adopt the worst case complexity in terms of the big-O notation. Based on this, the complexity of the steps of Algorithm 5.1 can be calculated as follows:

- steps 1–2 require $\mathcal{O}(U^2) + \mathcal{O}(K^2)$,
- steps 3–5 require $\mathcal{O}(K^2)$, and
- steps 6–21 require $\mathcal{O}(UK)$.

Therefore, the dominant component is $\mathcal{O}(U^2) + \mathcal{O}(K^2)$ which is that of constructing the preference lists of the users and PRBs because it contains sorting operations. This means that the computational complexity of the proposed HTC scheduling scheme in Algorithm 5.1 is much lower than that of the optimal solution of the combinatorial BLP problem in (5.12). This makes the algorithm more suitable in practice and real-time operation. Moreover, as will be analyzed in Section 5.7, the algorithm achieves a close-to-optimal performance, which shows that the reduction of the complexity does not noticeably degrade the quality of the solution as found by the algorithm.

5.6 Matching-Based Scheduling for the cMTC Traffic

For a complete practical scheduling scheme for all types of traffic, in this section, we propose a matching-based scheduling technique for the cMTC traffic and analyze it.

5.6.1 Matching Setup and Algorithm

Given the current matching of the HTC users and PRBs, μ^{UK}, we are trying to formulate the unified problem of scheduling cMTC traffic formulated in (5.46) as a two-sided matching process similar to what we accomplished in Section 5.5.1. In this case, the matching agents are the disjoint sets $\mathcal{M}$ and $\mathcal{K}$ that have preference lists over the opposite agents, $\mathcal{P}^M(m)$, and $\mathcal{P}^K(k)$, respectively. Therefore, the two-sided matching for this problem can be defined as follows.

Definition 5.6.1. *The matching assignment of the resource allocation problem in (5.46), μ^{MK}, can be defined as a mapping from the set $\mathcal{M} \cup \mathcal{K}$ into the set $\mathcal{M} \cup \mathcal{K}$ such that for any device $m \in \mathcal{M}$ and PRB $k \in \mathcal{K}$*

(i) $\mu^{MK}(m) \subseteq \mathcal{K}$,

(ii) $\mu^{MK}(k) \in \mathcal{M}$,

(iii) $|\mu^{MK}(k)| \leq 1$,

(iv) $|\mu^{MK}(m)| \geq q_m^{min}$,

(v) $k \in \mu^{MK}(m)$ *if and only if* $\mu^{MK}(k) = m$.

Conditions (i), (ii), and (iii) restrict the cardinality of the set of agents to which every type of agents can be matched to. Condition (iv) is used for the minimum quota bounds.

Similar to the matching process of the HTC scheduling, this process is a one-to-many matching problem. However, the new matchings of the PRBs in μ^{MK} should not violate the QoS requirements of the HTC users of the old match, μ^{UK}. This is ensured by applying conditions on the approval of any new matching for the PRBs. We propose the use of Algorithm 5.2 to solve such type of matching problems and we analyze it in the following.

5.6.2 Matching Analysis

In the same way as in Section 5.5.2, the convergence of the matching process described in Algorithm 5.2 can be proved as follows.

Lemma 5.6.2. *The matching process in Algorithm 5.2,* μ^{MK}*, converges to a final matching after a finite number of iterations.*

Proof. In Algorithm 5.2, every PRB $k \in \mathcal{P}^{comM}$ proposes to the most preferred device $m \in \mathcal{P}^{comK}$ until all devices are satisfied or the PRBs set $\mathcal{P}^{comM}$ is empty. Therefore, since $\mathcal{P}^{comM}$ is finite, the number of proposals is limited. Consequently, the matching algorithm converges to a final matching after a finite number of iterations. □

Algorithm 5.2 Matching-Based cMTC Scheduling Algorithm

Step 1: Initial setup

Calculate K_m^{min} for every $m \in \mathcal{M}$ such that its reliability constraint in (5.38) or (5.16a) be satisfied.

Construct the common preference list $\mathcal{P}^{comK}$ of all $k \in \mathcal{K}$ over $m \in \mathcal{M}$ in a descending order based on the required number of PRBs of the devices.

Construct the common preference list $\mathcal{P}^{comM}$ of all $m \in \mathcal{M}$ over $k \in \mathcal{K}$ based on the rate losses on every PRB, $R_{\mu^{UK}(k),k}$.

Step 2: Matching process

Set the status of every $m \in \mathcal{M}$ to 1, where a status of 1 indicates that the device is willing to propose, and otherwise is status 0.

while any device's status is 1 **do**

Feasibility test

  if $\mathcal{P}^{comM}$ is empty **then**

    Problem is infeasible. **Stop**

  end if

  $m^{propose} \leftarrow$ first m in $\mathcal{P}^{comK}$.

  $k^{prefered} \leftarrow$ most preferred k in $\mathcal{P}^{comM}$.

  Remove $k^{prefered}$ from preference list $\mathcal{P}^{comM}$.

  if $\mu^{UK}(k^{prefered})$ approves the unmatch based on its current rate **then**

    Match $m^{propose}$ to $k^{prefered}$.

    Update $\mu^{MK}(m^{propose})$, $\mu^{MK}(k^{prefered})$, $\mu^{UK}(k^{prefered})$, and $\varrho_{k^{prefered}}$.

    if $m^{propose}$ is now satisfied **then**

      Set its status to 0 and remove it from $\mathcal{P}^{comK}$.

    end if

  end if

end while

A blocking pair for this type of matching can be defined as follows.

Definition 5.6.3. *For a matching μ^{MK} as in Definition 5.6.1, a pair (m', k') is called a blocking pair if:*

1. $\mu^{MK}(k') \neq m',\ k' \notin \mu^{MK}(u')$,

2. $m' \succ_{k'} \mu^{MK}(k')$,

3. $k' \succ_{m'} k'',\ k'' \in \mu^{MK}(m')$, *and,*

4. *the lower quota bounds of m' and $\mu^{MK}(k')$ would still be fulfilled if k' is matched*

to m'.

Therefore, the stability of the matching scheme in Algorithm 5.2 can be proved as follows.

Theorem 5.6.4. *The matching process in Algorithm 5.2 admits no blocking pair and, hence, is stable.*

Proof. Algorithm 5.2 is designed such that every m is matched to the most preferred set of k and vice versa. This is achieved by exploiting the common preferences of the agents of the same type and ordering the preferred agents in two common lists, $\mathcal{P}^{comM}$ and $\mathcal{P}^{comK}$. Therefore, the first m in $\mathcal{P}^{comM}$, the most preferred by PRBs, proposes to the first k in $\mathcal{P}^{comK}$, the most preferred by devices, and so on. Accordingly, if (m', k') is a pair that satisfies $\mu^{MK}(k') \neq m',\ k' \notin \mu^{MK}(m'),\ k' \succ_{m'} k'',\ k'' \in \mu^{MK}(m')$ and the quota bounds of m' and $\mu^{MK}(k')$ would still be satisfied if m' is matched to k'; then, $m' \not\succ_{k'} \mu^{MK}(k')$. This is because if $m' \succ_{k'} \mu^{MK}(k')$, then it would propose to k' before $\mu^{MK}(k')$ and be matched to it since $\mu^{UK}(k')$ approves the unmatch of this PRB. Therefore, according to Definition 5.6.3, this pair cannot be a blocking pair. Accordingly, Algorithm 5.2 admits no blocking pair and is stable. □

In terms of the computational complexity, Algorithm 5.2 can be analyzed as follows. The elaborated complexity for the steps can be derived as:

- Steps 1–2 require $\mathcal{O}(M^2) + \mathcal{O}(K^2)$
- Step 3 requires $\mathcal{O}(K^2)$
- Steps 4–19 require $\mathcal{O}(M) + \mathcal{O}(K)$

Accordingly, $\mathcal{O}(M^2)+\mathcal{O}(K^2)$ is the worst case computational complexity of Algorithm 5.2. Therefore, the computational complexity of the algorithm for the scheduling of

Table 5.2: Simulation Parameters of Chapter 5

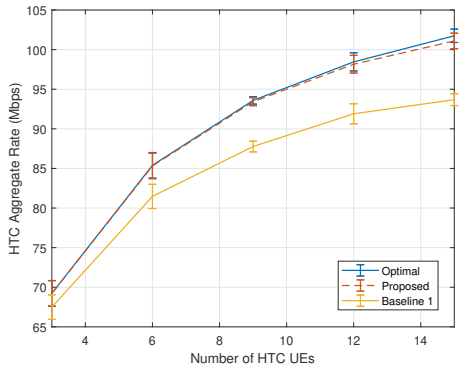
Parameter	**Value**
Cell radius	500 m
Path loss L (dB)	$128.1 + 37.6 \log_{10}(d)$, d in km [24]
Standard deviation of shadowing (σ)	8 dB
eNB transmit power (P)	40 dBm
Power spectral density of noise	-174 dBm/Hz
Noise figure	18 dB
Bandwidth	10 MHz
HTC average arrival rate	$128, 256, 512$ Kbps
cMTC average arrival rate	$20, 30, 40, 50, 60$ Kbps
cMTC packet size	160 bits
Delay bound (D_m^{max})	0.3 ms
Maximum PDBV (ε_m)	10^{-3}
Maximum PTE (ϵ_m^{max})	10^{-5}
$\bar{R}_u^{min},\ u \in \mathcal{U}$	λ_u

the other type of traffic indicates that it can be used in practice with a reduced complexity.

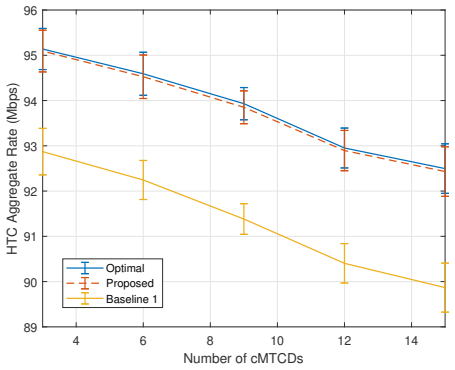
5.7 Simulations Results

In this section, we present and discuss the results of the simulations conducted to evaluate the performance of the proposed matching-based scheduling algorithms in an LTE system with practical parameters. In addition, we compare the performance of the proposed methods with that of the optimal solution of the formulated scheduling problems in (5.12) and (5.46) to prove how close the performance of the proposed algorithms is to the optimal benchmark. Moreover, we illustrate the superiority of the performance compared to baseline scheduling algorithms.

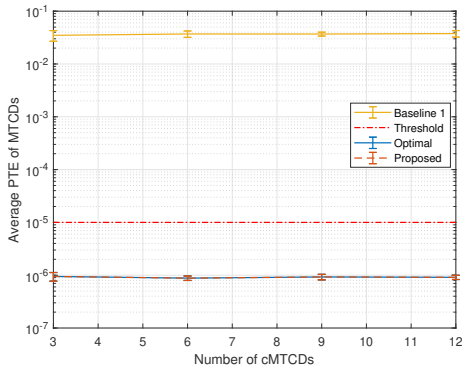
We have conducted two simulations for the considered two cases on MATLAB. In both simulations, we consider a single LTE cell of radius 500 m that contains



(a) HTC sum-rate vs no. of HTC UEs ($M = 12$).



(b) HTC sum-rate vs no. of cMTCDs ($H = 9$).

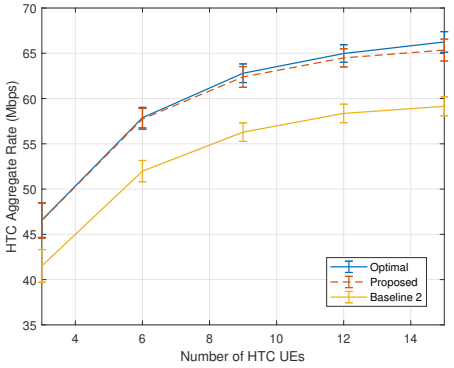


(c) cMTC average PTE vs no. of cMTCDs ($H = 9$).

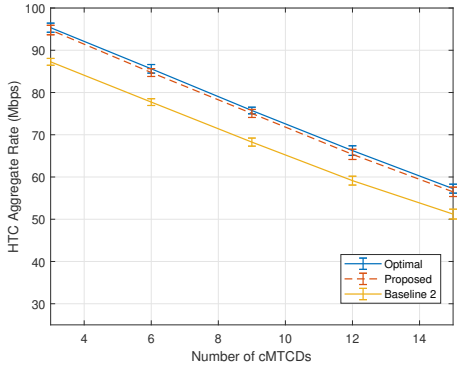
Fig. 5.3: Performance evaluation and comparison for Case 1.

a single eNB at the center. The eNB serves two sets of users, i.e., HTC UEs and cMTCDs, that are uniformly distributed within the cell. The path loss in dB is calculated by $128.1 + 37.6 \log_{10}(d)$, where d (in km) is the distance between the user and the eNB [24]. The shadowing is assumed to be log-normal with a standard deviation of 8 dB and the fading is i.i.d. complex Gaussian. We assume that the traffic arrivals follow Poisson distribution with average arrival rates as shown in Table 5.2, which summarizes the used values of the parameters of the simulations. The QoS requirements of the cMTCDs are expressed as latency bounds, D_m^{max}, and reliability guarantees in terms of maximum PDBV and PTE, with values as in Table 5.2. On the other hand, we ensure an average rate of every HTC UE that is at least equal to its average arrival rate. Those two types of traffic are scheduled on two different scales of time, i.e., TTIs and sTTIs, as discussed in Section 5.3.

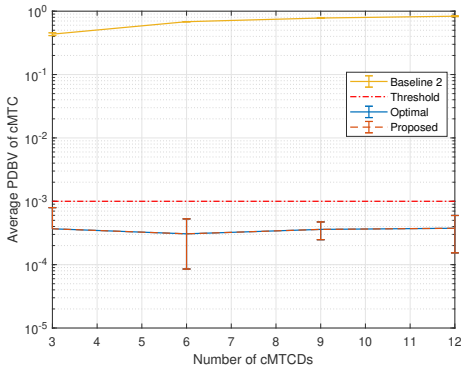
For comparison purposes, we calculate the optimal solutions of the optimization problems in (5.12) and (5.46) for the scheduling of the HTC and cMTC, respectively. For this purpose, we use MATLAB's Optimization Toolbox. The baseline scheduling algorithm for Case 1 uses the proportional fairness (PF) scheduler for the HTC traffic at every TTI. Then, the cMTC traffic punctures the HTC transmissions by calculating the required number of PRBs, based on Shannon capacity, such that the transport block size (TBS) fits the cMTC packet. The PRBs to be punctured are selected in a round robin manner, which is similar to the approach used in [16] that selects the PRBs randomly. On the other hand, the baseline scheduler of Case 2 adopts PF scheduling algorithm for both types of traffic and uses the same bandwidth for cMTC as that of the optimal scheduler. The adopted performance indicators are the HTC aggregate data rate and the average PTE and PDBV of cMTC. We plot the confidence intervals as bars on every point to show the accuracy of estimating the metric in the simulations.



(a) HTC sum-rate vs no. of HTC UEs ($M = 12$).



(b) HTC sum-rate vs no. of cMTCDs ($H = 15$).



(c) cMTC average PDBV vs no. of cMTCDs ($H = 9$).

Fig. 5.4: Performance evaluation and comparison for Case 2.

Figure 5.3 shows the performance evaluation and comparison for Case 1. In Figs. 5.3(a) and 5.3(b), the aggregate data rate of the HTC traffic, achieved using the three different scheduling algorithms, is plotted versus the number of HTC UEs and cMTCDs, respectively. As expected, the HTC sum-rate improves with the increase of the number of HTC UEs and degrades with increasing the number of served cMTCDs in the cell. The latter is due to the fact that serving more cMTCDs entails more puncturing, which results in more HTC data rate loss. In both figures, the matching-based scheduler achieves a close-to-optimal HTC sum-rate with a gap that increases with the increase of the number of users, which is expected as a sub-optimal solution with lower complexity. In addition, the figures show how the optimal and proposed algorithms outperform the baseline scheduler. This is the result of not optimizing the selection of the HTC PRBs to be punctured by the cMTC traffic, which degrades the sum-rate of the HTC traffic as a consequence. In Fig. 5.3(c), the average PTE of cMTC in the cell is plotted against the number of cMTCDs. Since the PTE constraint in (5.16a) is expressed as a minimum number of PRBs constraint as in (5.40), both the optimal and matching-based algorithms satisfy that constraint with equality. This is due to the fact that the more the HTC PRBs are punctured, the more the HTC rate loss results. Therefore, both algorithms satisfy the number of PRBs constraint with equality. However, the selection of which HTC PRBs to be punctured is different. This yields a difference in the HTC sum-rate as in Figs. 5.3(a) and 5.3(b). On the other side, the baseline scheduler calculates the data rate of cMTCDs based on Shannon capacity and does not consider the PTE due to the FBC. This results in a violation of the restricted PTE threshold.

Figure 5.4 depicts the calculated performance metrics for Case 2. Figs. 5.4(a) and 5.4(b) show the sum-rate of the HTC traffic versus the number of users and devices. In a similar behavior to Case 1, the proposed matching-based algorithms achieve

a close-to-optimal performance and outperforms the baseline scheduler, which does not target minimizing the data rate loss of the HTC UEs. In addition, the baseline scheduler violates the PDBV constraints of the cMTC traffic since it does not consider the queuing delay of the cMTC packets. However, both of the matching-based and optimal schedulers consider the PDBV of the cMTCDs as a constraint. Similar to the PTE in Case 1, the PDBV constraint in (5.38) is expressed as a minimum number of PRBs constraint as in (5.40). This makes both of the proposed and optimal solutions satisfy the constraint with equality and result in the same PDBV behavior.

5.8 Conclusions

In this chapter, we addressed the resource allocation problem for cMTC and HTC traffic in LTE networks that support sTTIs using a puncturing scheduling technique. To provide guarantees for the reliability of cMTC, we considered finite blocklength coding analysis to model transmission errors and effective bandwidth and effective capacity concepts for the queuing delay. This covers the two major sources of packets losses. The formulated problems targeted maximizing the data rate of the HTC traffic, minimizing their losses due to puncturing and satisfying the QoS requirements of all users and devices. Computationally-efficient matching-based scheduling schemes were proposed to solve the formulated problems efficiently with much lower complexity. The proposed methods were analyzed from a practical point of view. The simulation results showed a close-to-optimal performance of the proposed schemes and its superiority to baseline scheduling algorithms.

References

[1] ITU-R M.2083, *IMT vision - Framework and overall objectives of the future development of IMT for 2020 and beyond*, Sept. 2015.

[2] 3GPP TR 38.913, *Study on scenarios and requirements for next generation access technologies, technical specification group radio access network*, Oct. 2016.

[3] 3GPP RP-171489, *Ultra Reliable Low Latency Communication for LTE*, June 2016.

[4] 3GPP RP-161299, *Work Item on shortened TTI and processing time for LTE*, June 2016.

[5] 3GPP RP-160667, *Work item on L2 latency reduction techniques for LTE*, March 2016.

[6] C. Sun, C. She, C. Yang, T. Q. Quek, Y. Li, and B. Vucetic, "Optimizing resource allocation in the short blocklength regime for ultra-reliable and low-latency communications," *IEEE Transactions on Wireless Communications*, vol. 18, no. 1, pp. 402–415, 2019.

[7] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Transactions on Information Theory*, vol. 56, no. 5, p. 2307, 2010.

[8] L. You, Q. Liao, N. Pappas, and D. Yuan, "Resource optimization with flexible numerology and frame structure for heterogeneous services," *IEEE Communications Letters*, vol. 22, no. 12, pp. 2579–2582, 2018.

[9] 3GPP TSG RAN WG1 Meeting 87, November 2016.

[10] K. I. Pedersen, G. Pocovi, J. Steiner, and S. R. Khosravirad, "Punctured scheduling for critical low latency data on a shared channel with mobile broadband," in *2017 IEEE 86th Vehicular Technology Conference (VTC-Fall)*. IEEE, 2017, pp. 1–6.

[11] C. She, C. Yang, and T. Q. Quek, "Cross-layer optimization for ultra-reliable and low-latency radio access networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 1, pp. 127–141, 2018.

[12] ——, "Joint uplink and downlink resource configuration for ultra-reliable and low-latency communications," *IEEE Transactions on Communications*, vol. 66, no. 5, pp. 2266–2280, 2018.

[13] A. Anand and G. de Veciana, "Resource allocation and harq optimization for urllc traffic in 5g wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 11, pp. 2411–2421, 2018.

[14] M. Y. Abdelsadek, Y. Gadallah, and M. H. Ahmed, "Optimal cross-layer resource allocation for critical mtc traffic in mixed lte networks," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 6, pp. 5944–5956, June 2019.

[15] ——, "An lte-based optimal resource allocation scheme for delay-sensitive m2m deployments coexistent with h2h users," in *2017 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, May 2017, pp. 139–144.

[16] A. Anand, G. De Veciana, and S. Shakkottai, "Joint scheduling of urllc and embb traffic in 5g wireless networks," in *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*. IEEE, 2018, pp. 1970–1978.

[17] M. Alsenwi, N. H. Tran, M. Bennis, A. K. Bairagi, and C. S. Hong, "embb-urllc resource slicing: A risk-sensitive approach," *arXiv preprint arXiv:1902.01648*, 2019.

[18] S. Schiessl, J. Gross, and H. Al-Zubaidy, "Delay analysis for wireless fading channels with finite blocklength channel coding," in *Proceedings of the 18th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*. ACM, 2015, pp. 13–22.

[19] C.-S. Chang and J. A. Thomas, "Effective bandwidth in high-speed digital networks," *IEEE Journal on Selected areas in Communications*, vol. 13, no. 6, pp. 1091–1100, 1995.

[20] D. Wu and R. Negi, "Effective capacity: a wireless link model for support of quality of service," *IEEE Transactions on wireless communications*, vol. 2, no. 4, pp. 630–643, 2003.

[21] M. David, *Algorithmics Of Matching Under Preferences*. World Scientific, 2013, vol. 2.

[22] F. Kelly, S. Zachary, and I. Ziedins, *Stochastic Networks: Theory and Applications*. Oxford University Press, 1996.

[23] K. Hamada, K. Iwama, and S. Miyazaki, "The hopitals/residents problem with quota lower bounds," in *MATCH-UP 2008: Matching Under Preferences, satellite workshop of ICALP 2008*. IEEE, 2008, pp. 55–66.

[24] *Further advancements for E-UTRA physical layer aspects*, 3GPP TR 36.814, 2010.

Chapter 6

Conclusions and Future Work

In this chapter, we conclude this dissertation, in Section 6.1. Then, we present some possible extensions of this work in Section 6.2.

6.1 Conclusions

In this dissertation, we designed resource allocation and scheduling schemes to efficiently support critical MTC traffic in LTE networks without degrading the QoS of conventional HTC, nor the overall system throughput. Due to the fact that the data transmissions of critical MTC are characterized by their low data rate and small packet size, in contrast to the data-hungry HTC applications, we formulated the resource allocation problem such that the aggregate data rate of the HTC is maximized while providing guarantees for the QoS requirements of both types of traffic. In this manner, we formulated the resource allocation problem while considering different operational cases and techniques.

Simulations proved that our problems' formulations yield the best cell throughput compared to other scheduling techniques, such as the well-known proportional fairness scheduler. In addition, the simulations revealed the satisfaction of the QoS

demands of all users and devices in all cases, since they are considered as constraints to the optimization problem. Accordingly, any feasible solution must satisfy these constraints.

To provide QoS guarantees for critical MTC so that designers can build their applications based on concrete platforms of wireless connectivity, we used exact models from the queueing theory for some scenarios and accurate approximation models from the effective bandwidth and effective capacity theories for the others. The simulations validated the designed models and proved that the buffer dynamics of the devices well-fit the used models.

In all of the considered scenarios of system model, we derived the optimal solution of the formulated optimization problem. However, the optimal algorithms cannot be implemented in practice as scheduling schemes due to their high computational complexity. Therefore, for every system model, we proposed a practical, low-complexity scheduling scheme that can solve the formulated optimization problem, satisfying the constraints, but with sub-optimal or local optimal objective value. The proposed algorithms utilized heuristic approaches and the matching theory. The simulations proved that the performance of the proposed algorithms is still close to the optimal solution and, at the same time, better than other scheduling techniques from the literature. On the other hand, the computational complexity has been proved to be much lower than that of the optimal schemes. This is in addition to its guaranteed stability and convergence.

As indicated above, it can be concluded that critical MTC can be efficiently served in LTE networks without degrading the QoS of the conventional HTC traffic, from the radio resource management perspective. This can be achieved by formulating the resource allocation problem such that the QoS requirements of critical MTC and that of HTC are considered as constraints. The objective function to be maximized is the

aggregate data rate of the HTC traffic, without maximizing that of the critical MTC. For accurate guarantees for the satisfaction of the QoS requirements of critical MTC, the effective bandwidth and effective capacity theories can be used in the general scenario. Besides, exact models from the queueing theory can be utilized in some special cases.

In addition, the newly introduced techniques and procedures in LTE can be used to serve critical MTC more efficiently. The additional degrees of freedom introduced by massive MIMO can be exploited to serve more critical MTC devices, minimize the impact on HTC, and achieve higher throughput of the LTE cell. Besides, puncturing scheduling technique can be used, taking advantage of the support of the short transmission intervals in LTE, to minimize the data rate loss of HTC and support very low latency MTC applications.

Moreover, for practical implementation considerations, matching-based scheduling algorithms can be used in the real-time process. They exhibit superior performance while being implemented in polynomial time.

6.2 Future Work

Although this dissertation considers several aspects of the design of the resource allocation and scheduling for critical MTC in LTE networks, there are some possible extensions to the current work such as:

- Although the matching-based resource allocation schemes are efficient algorithms in terms of performance and complexity, machine-learning represents a good candidate for another approach to solve the problem. Machine-learning has received much attention in the literature and has mature techniques and methodologies that can be employed in the radio resource management area.

The current work can be utilized in the machine-learning approaches to a large extent since the problem formulation and cross-layer design are already developed.

- In addition, this work can be extended by taking into consideration the coordination between the cells. In this regard, coordinated multi-point (CoMP) is a technique that is used in LTE to coordinate the transmission and reception for a user equipment using several LTE eNBs. Therefore, the proposed algorithms can be extended to the multiple-cells scenario.

- Another potential extension is to consider the procedures and numerology of the 5G new radio (NR). This is due to the fact that the NR adopts similar procedures as the LTE. Besides, techniques such as non-orthogonal multiple access (NOMA) can be utilized to optimize the support of the critical MTC in 5G cellular networks.