QATAR UNIVERSITY

COLLEGE OF ENGINEERING

DYNAMIC RESOURCE ALLOCATION OF EMBB/URLLC TRAFFIC IN 5G NR

BY

MUHAMMED AL-ALI

A Thesis Submitted to
the College of Engineering
in Partial Fulfillment of the Requirements for the Degree of
Master of Science in Computing

January 2021

COMMITTEE PAGE

The members of the Committee approve the Thesis of
Muhammed Al-Ali defended on 25/11/2020.

Dr. Elias Yaacoub
Thesis/Dissertation Supervisor

Prof. Mohsen Guizani
Committee Member

Prof. Amr Mohamed
Committee Member

Dr. Loay Ismail
College Representative

Approved:

Khalid Kamal Naji, Dean, College of Engineering

ABSTRACT

AL-ALI, MUHAMMED S., Masters : January : 2021,
Masters of Science in Computing

Title: DYNAMIC RESOURCE ALLOCATION OF EMBB/URLLC TRAFFIC IN 5G NR

Supervisor of Thesis: Elias, E, Yaacoub.

5G technology is intended to support three promising services with heterogeneous requirements: Ultra-Reliable and Low Latency Communication (uRLLC), enhanced Mobile Broadband (eMBB) and massive Machine Type Communication (mMTC). The presence of these services on the same network creates a challenging task of resource allocation to meet their diverse requirements. Given the critical nature of uRLLC applications, uRLLC traffic will always have the highest priority which causes a negative impact on the performance of other types of services.

In this thesis, the problem of uRLLC/eMBB resource allocation is addressed. Sub-optimal and optimal solutions are proposed. Heuristic scheduling algorithms are utilized in the sub-optimal approach, providing a low complexity solution to the problem. A knapsack inspired punctured resource allocation algorithm is proposed in which the channel quality of both eMBB and uRLLC UEs are considered at each time slot to make the best Resource Block (RB) selection for puncturing in a way that minimizes the impact on eMBB performance. In addition, the proposed algorithm is compared with three reference algorithms with similar objectives and the performance is evaluated in terms of eMBB Spectral Efficiency, Sum throughput and Fairness level.

The simulation results show that the proposed algorithm outperforms the above-mentioned reference algorithms in all evaluation metrics and showed its capability of elevating the performance of heuristic scheduling algorithms in the presence of uRLLC traffic.

In the second part of this thesis, an optimal resource allocation scheme with guaranteed fairness is proposed in which, it can provide the desired level of fairness among eMBB users while maximizing their data rate. The results show how the fairness level and the number of uRLLC users affect the performance of the algorithm using different intensities. It also shows that the optimal allocation scheme provides better results in terms of eMBB sum-throughput while preserving the desired level of fairness.

DEDICATION

To my beloved family

For their unlimited love and support.

ACKNOWLEDGMENTS

I thank Allah for giving me the strength to complete this journey and to accomplish this work. I would like to thank my family for their support and understanding throughout my two years of study. I also would like to express my sincere appreciation to Dr. Elias Yaacoub for his patience, guidance, and continuous support.

This work was made possible by NPRP grant # NPRP12S-0119-190006 from the Qatar National Research Fund (a member of Qatar Foundation). The findings achieved herein are solely the responsibility of the authors.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

uRLLC Ultra Reliable and Low Latency Communication

eMBB Enhanced Mobile Broadband

mMTC Massive Machine Type Communication

UE User Equipment

BS Base Station

RB Resource Block

3GPP the third Generation Partnership Project

VR Virtual Reality

AI Artificial intelligence

IoT Internet of Things

SC Sub-Carrier

IMT International Mobile Telecommunications

TTI Transmission Time Interval

BER Bit Error Rate

BLER Block Error Rate

ITS Intelligent Transportation systems

CQI Channel Quality Indicator

BW Bandwidth

SCS Sub-Carrier Spacing

OFDM Orthogonal Frequency Division Multiplexing

M-MIMO Massive – Multiple Input Multiple Output

UDN Ultra-Dense Networks

Gbps Giga bits per second

DC Dual Connectivity

LDPC Low density parity check

ICI Inter-Carrier Interference

CP Cyclic prefix

LTE Long Term Evolution

UL Uplink

DL Downlink

KHz Kilo Hertz

μs Microsecond

M-LWDF Maximum Weighted Delay First

PF Proportional Fairness

Ex-PF Exponential weighted Proportional Fairness

CHAPTER 1: INTRODUCTION

The massive technological development in electronic devices facilitated the emergence of new applications (e.g., Artificial intelligence (AI), Big Data analysis, Internet of everything, Virtual Reality (VR), etc.) having a ubiquitous influence on people's lives. Nonetheless, these applications produce a huge amount of data traffic in addition to requiring continuous connectivity, raising one of most challenging tasks for today's cellular communication technologies to overcome. As shown in Figure 1, its expected for Smart Phones, Tablets, Routers and Mobile PCs combined data traffic to reach 164 exabytes/Month by the end of 2025 which is 5 times the amount of traffic in 2019 [1]. By observing these figures, it is almost certain that the amount of traffic will continue to increase rapidly in the future which frames the technical objectives of the desired cellular system that includes the following:

- High data rates that could reach tens of Giga bits per second (Gbps) per device or even more on area basis. This is mainly caused by video traffic which is expected to grow by 30% annually occupying three-quarters of the total traffic by 2025 [2].
- Massive number of connected devices which by itself raises many challenges including interference among all these transmitting devices. This is due to the fact that users with multiple devices would require multiple connections in addition to the Internet of Things (IoT) which demands networks designed to handle billions of devices not to mention that the operational performance of the cellular system also depends on increasing the energy efficiency while serving all these devices [1][2].
- Extremely low latency required by applications such as VR and Remote surgery which is solely based on tactile interaction [2].

- Ultra-reliability level which is mandatory to support applications with a certain level of criticality (e.g. Remote surgery, Intelligent Transportation systems (ITS), High Voltage Electricity distribution and Industrial Control).

Among all these factors its worth mentioning that network operators are forced to reduce the operational expenditure to meet the demands of the users which includes stable and flat data rates with low service prices leading to unwanted restriction to system designers forcing them to consider affordable designs and limit the use of the technologies to its full extent.

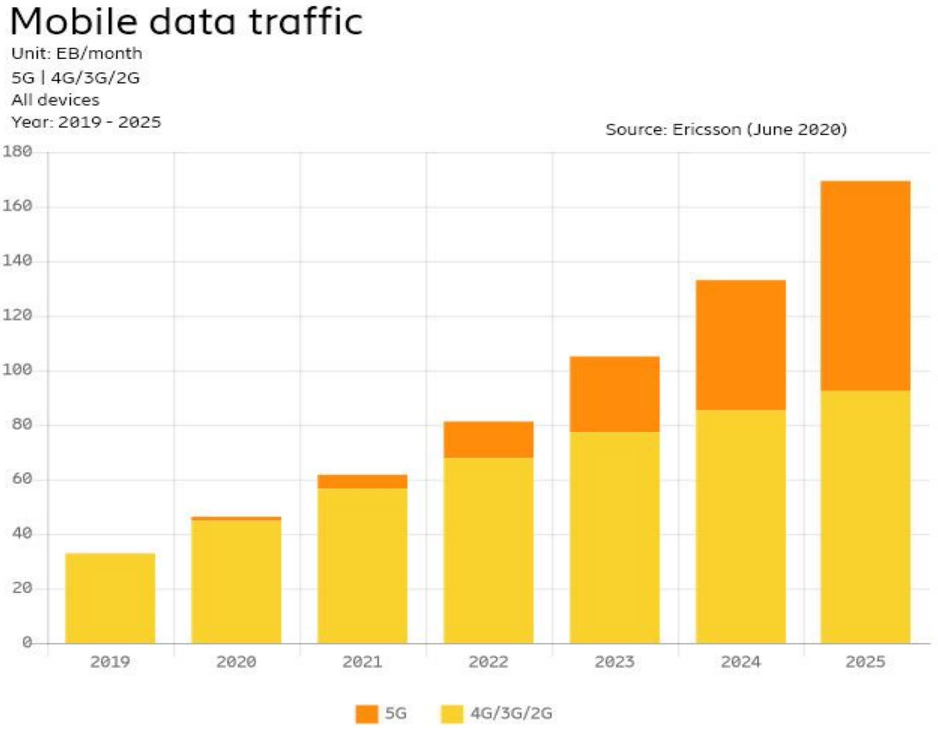

Figure 1. Mobile Data Traffic Estimations [1]

As a result, the currently used Fifth Generation (5G) cellular network is pushed to provide high operational performance including high data rates, spectral efficiency, low latency, and high reliability aiming to fulfill the user’s experience demands. This requires an acceptable equipment cost, power consumption level and network operational costs. Figure 2 demonstrates the enhancements from the Fourth-Generation cellular system (4G) (IMT-advanced) to 5G (IMT-2020) in terms of capabilities aiming to cope with the above-mentioned requirements caused by the explosive growth of data traffic volumes [2].

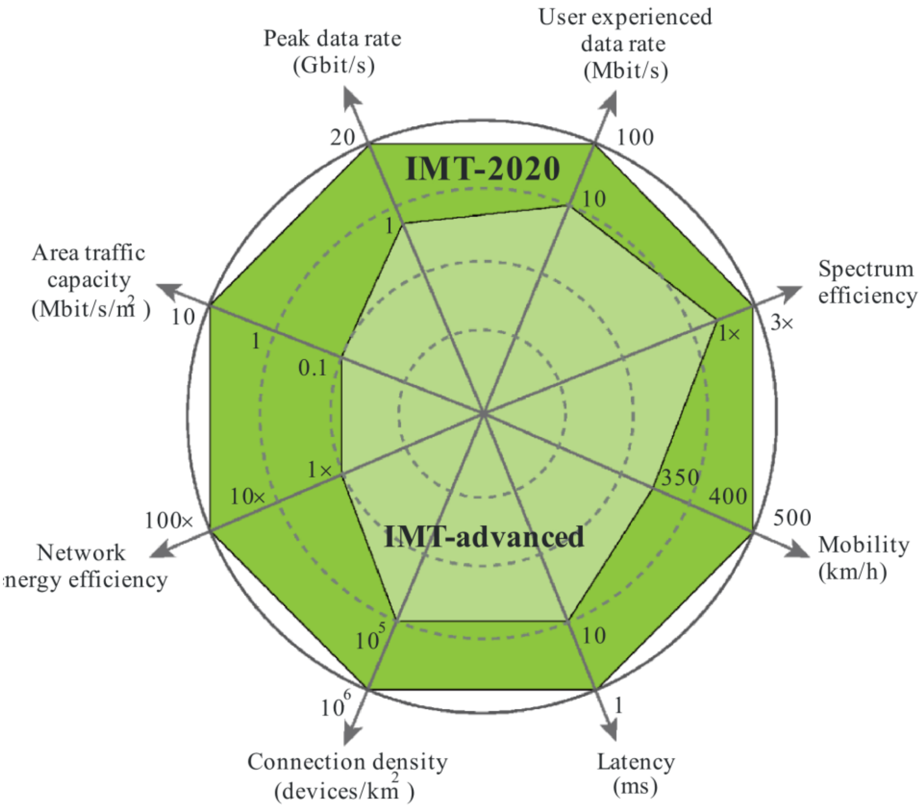

Figure 2: Enhancement of key capabilities from IMT-Advanced to IMT-2020 [3]

Handling these heterogeneous requirements is a challenging task, and it is considered as one of the reasons the International Telecommunication Union (ITU) classified the services 5G is envisioned to support into different categories [3].

1.1 5G Service Categories

1.1.1 enhanced Mobile Broadband (eMBB).

This communication service is an extension of the traditional Mobile Broadband used in LTE. It aims to provide high data rates, high user mobility and better connectivity which is essential for human-centric applications such as online gaming and ultra-high-quality video streaming. The objective of offering eMBB in 5G networks was to achieve a peak data rate of 20 Gbps and a moderate reliability level of 10^{-3} Packet loss rate. Some of the eMBB related scenarios include Hotspot connectivity that features large user densities with low mobility and high data rates [3][4].

1.1.2 massive Machine Type Communications (mMTC).

mMTC is also a communication service designed to provide an efficient connectivity to a massive number of devices and it is considered as one of the main enablers of Internet of Things. mMTC is involved in numerous applications such as traffic and environmental monitoring, resource utility management and smart grids. In addition to massive connectivity, it’s important to mention that the devices in which mMTC deals with are battery-powered and tend to stay running for a long period of time (e.g. 1-2 years) and thus, high coverage and extreme indoor penetration is essential for these devices to operate efficiently [5].

1.1.3 ultra-Reliable and Low Latency Communication (uRLLC).

uRLLC is a communication service specifically designed for mission critical applications, this includes industrial automation, electricity distribution, tactile interaction, and intelligent transportation service. The main features of uRLLC are defined with its strict latency, reliability, and availability requirements [6].

uRLLC was intended to deal with relatively small sized payload of about 256 bits while maintaining a user plane latency of 1 ms, control plane latency of 10 ms and a reliability level of 99.999 % aiming to transfer the data to the upper layers in a fast and error-free form while preserving it's integrity, confidentiality and authenticity [7][8]. Figure 3 shows the key features of the above mentioned 5G use cases.

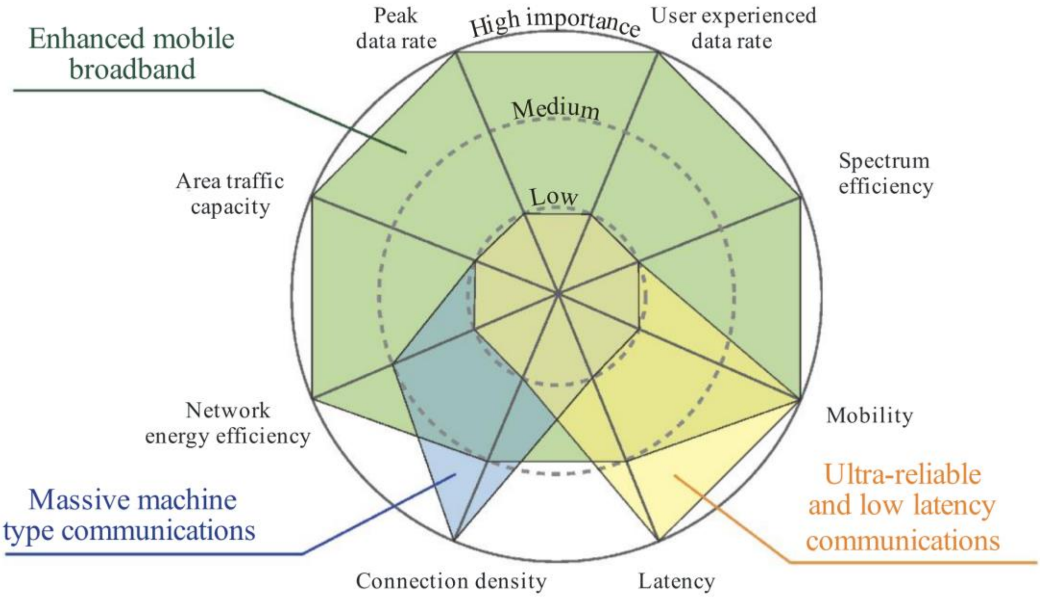

Figure 3: Capabilities of 5G use cases. [3]

While the main design concept of the Fourth Generation (4G) cellular system was to support human-centric applications such as Mobile Broadband (MBB) and voice/video communication by providing high-speed connectivity for a relatively limited number of users, the fifth generation (5G) cellular system was designed to overcome the limitations of 4G in addition to supporting machine-centric applications that are known with their massive numbers and diverse requirements.

1.2 Motivation and Problem Statement

While classifying the services into different classes helps in identifying the applications in which these services are used by and thus assigning suitable priority level for each in order to service all applications efficiently, this created a new obstacle towards achieving the best operational performance. This issue is the coexistence of these services with their heterogeneous requirements within the same network infrastructure and given that the operators are tied by a finite bandwidth (BW) and limited operational cost budget, cellular system design constraints are created for the operators. As a result, this issue is considered as an interesting research topic addressed by both the industry and the academia.

What triggers this interest is the great potential these services have and how promising their contribution is envisioned as part of the future cellular communication systems given the diversity of applications, they are capable of serving in different industries. Although all 5G main use cases have their uniqueness and promising potential for the current and future applications, uRLLC captured the focus of researchers more than ever recently. Applications in Healthcare, Transportation and Tactile interaction heavily rely on uRLLC to operate and we can notice that some of these applications have a direct impact on our lives.

What is envisioned in these three domains is life changing, from remote patient diagnosis and surgeries to self-driving cars and fully operating smart cities. Nevertheless, one of the most challenging tasks in enabling uRLLC is the existence of eMBB in the same network infrastructure which creates a difficult resource allocation task for network operators in order to satisfy their Quality of Service (QoS) requirements.

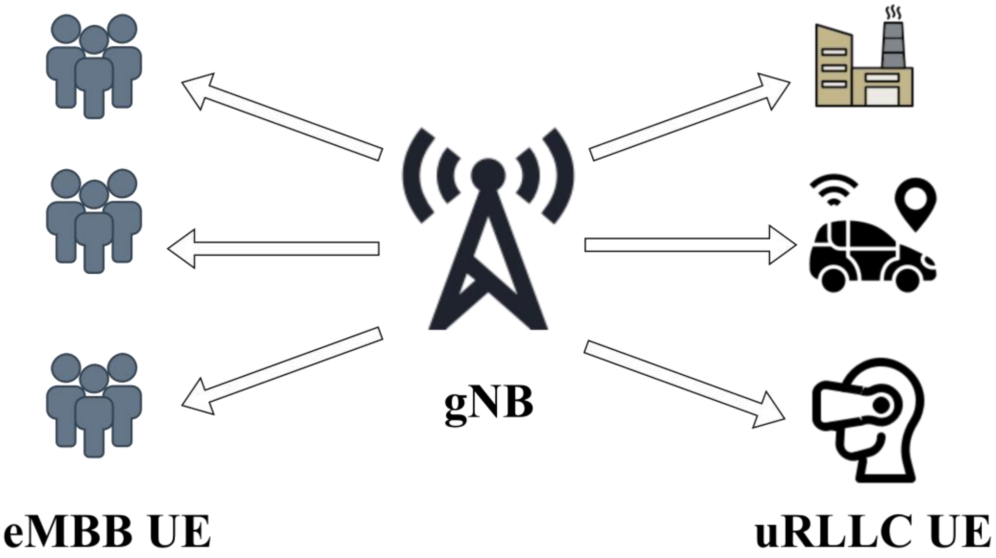

Figure 4: eMBB/uRLLC coexistence

The limited amount of resources is not the only issue in the presence of uRLLC with other services, but the fact that the stochastic nature of uRLLC makes it even more difficult to deal with its unexpected arrival at the BS. As mentioned above, the requirements of uRLLC forces the BS to serve the incoming traffic within one Transmission Time Interval (TTI) (1 ms). Nevertheless, the payload size, the availability of resources and the unstable radio channel condition might force the BS to schedule the incoming traffic to the next time slot.

The 3rd Generation Partnership Project (3GPP) proposed two scheduling approaches to handle the uRLLC traffic. The first approach is known as reservation-based scheduling while the other one is known as instant scheduling (Preemptive/Puncturing scheduling) [29][30][31]. The first approach uses a uRLLC reservation-based frame to handle any unexpected traffic. It can either use static or dynamic resource reservation. Static reservation method tends to send the frame structure that holds the transmission configurations (e.g. adapted Numerology) in an intermittent fashion.

Unlike the static reservation method, in dynamic reservation, the frame structure is sent frequently to the UE. This approach causes a control signaling overhead and the resources reserved to the uRLLC might be wasted in the case where there are no incoming uRLLC data.

The second approach (Known as Instant scheduling) aims to serve any incoming uRLLC traffic instantly using short TTIs of 2,4,7 OFDM symbols (Mini slot-based scheduling) . While this approach might cause an interruption to ongoing transmissions of other applications and might cause a huge performance degradation of other services like mMTC and eMBB, it is still considered a more efficient approach as it can be relied on to support the strict latency requirements of uRLLC.

Several techniques have been used in the literature and can be classified into two categories, the first one is the use of Network Slicing in which the network is subdivided into several logical networks known as network slices. Each slice is configured according to the requirements of the service operating over it. Network slicing can be supported by Software Defined Networks providing more flexible design [1]. Other approaches are based on Resource Puncturing/Superposition in which pre-allocated resources to eMBB UEs are given to uRLLC traffic upon arrival.

Both techniques formulate the resource allocation problem as an optimization problem in which they tend to solve it using different methods including machine learning, game theory and convex optimization algorithms. Nevertheless, the efficiency of these optimal allocation schemes is questionable when implemented in practice because of their high complexity in addition to the stochastic nature of the uRLLC traffic that make this task even harder. As a result, a few papers in the literature propose a sub-optimal allocation algorithm in the coexistence of eMBB and uRLLC traffic in which the main objective is to serve them both efficiently and with low complexity [1].

This can make these approaches more practical and perhaps more favorable over optimal allocation schemes. Below are the basic research questions we aim to answer in this thesis.

1. How to degrade the impact of uRLLC on other types of services like eMBB ?
2. What are the most suitable resource allocation algorithms to deal with the coexistence of eMBB and uRLLC in 5G network ?
3. What are the key features of these algorithms and how do these features contribute to satisfying the requirements of both services efficiently ?
4. What are the tradeoffs between maximizing the throughput and maintaining fairness among users ?
5. How to utilize heuristic scheduling algorithm and benefit from their low complexity in the presence of eMBB and uRLLC services in the same network ?
6. What kind of improvements can be done to these schedulers in order to make them more suitable for use in such environment ?
7. How significant is the difference between optimal and sub-optimal resource allocation of uRLLC and eMBB traffic ?

1.3 Thesis Objectives

The main objective of this thesis study is to provide the most suitable resource allocation strategy of uRLLC/eMBB traffic in 5G network. The aim is to lower the impact of uRLLC traffic on eMBB UE while satisfying the strict latency requirements of uRLLC (1 ms).

A sub-optimal approach has been adopted by using heuristic scheduling algorithms which are known with their low complexity and might be considered a more efficient solution to the resource allocation problem taken into consideration the high complexity of optimal allocation schemes.

Four resource puncturing algorithms are proposed aiming to find the most suitable RB at every time slot for puncturing according to different criteria. The sub-optimal approach includes four objectives with a common constraint of satisfying the requirements of the existing uRLLC traffic in each time slot. The objectives are listed below:

1. Providing the best resource block for uRLLC UE where they experience the best channel conditions. This aim is to allocate these RBs in order to minimize the amount of punctured resources and thus affecting the data rates of eMBB UE. Allocating RBs for uRLLC UE with better channel condition would lead to a better spectral efficiency for uRLLC (i.e. more data are sent using this RB when compared to other RBs).
2. The second objective is to protect eMBB UE at the cell edge where these users most likely suffer from bad channel conditions and thus cannot tolerate puncturing their resources. This can be achieved by targeting eMBB UE with better channel conditions by allocating their resources for uRLLC traffic.

3. The third objective is to maximize the sum rate of eMBB UEs. This can be achieved by targeting eMBB UE with low channel conditions and thus have less contribution to the overall sum throughput of eMBB UEs.
4. The four objective is to maximize eMBB sum throughput while maintaining an acceptable level of fairness among the users in terms of the amount of punctured resources from each user. This can be achieved using a knapsack inspired resource allocation scheme that considers the channel quality of uRLLC UE in addition to the achievable data rate of eMBB UE at each RB when making the puncturing decision (i.e. selection of RB at time slot t for puncturing).

1.4 Thesis Overview

This thesis is organized as follows.

Chapter 1 provides an introduction about the topic in addition to related background information about 5G technology and the different services it supports. It also includes the problem statements and the main objectives along with proposed approaches to achieve each objective.

Chapter 2 includes an overview about uRLLC, its applications and enablers reflecting its importance. In addition, an overview about heuristic scheduling algorithms, explaining their utility functions, input parameters and key features is presented. Also, a literature review of the state-of-the-art works addressing the resource allocation problem of uRLLC and eMBB traffic and the different approaches proposed in these papers are explained.

In Chapter 3, a sub-optimal, puncturing based resource allocation approach is proposed utilizing heuristic scheduling algorithms in existence of uRLLC and eMBB traffic in the same network.

Chapter 4 presents an optimal allocation of uRLLC/eMBB traffic that preserves the desired fairness level while considering the stochastic nature of uRLLC.

Chapter 5 includes the performance evaluation of the proposed approaches including numerical results of our simulations.

Chapter 6 presents the research findings, some limitations and possible future work for improvement purposes.

CHAPTER 2: BACKGROUND / RELATED WORK

In this section, a brief literature review is conducted in addition to some background study on some of the technologies and concepts used in this thesis.

As the main topic of this thesis is about uRLLC, it is important to study some of its applications and their different requirements in order to realize the critical nature of these applications and to justify their strict requirements. These applications give us an idea of the different scenarios in which the base station might encounter and how any resource allocation decision is affected by several parameters including different latency requirements and traffic densities.

2.1 Applications of Ultra-Reliable and Low Latency Communication [9][10]

Wireless technologies are considered as major enablers of today’s digitalized, data-driven, and hyper connected society. This section provides an overview about uRLLC and reflects its importance in enabling many vital applications used in different domains. Enabling uRLLC is one of the most primary objectives of 5G as it is involved in several mission-critical applications in which some of them are life-changing and might be considered revolutionary given their huge impact on people’s lives. Some of these applications are listed below along with their importance and requirements.

2.1.1 Medical and Healthcare [12][13]

uRLLC plays a crucial role in enabling one of the most revolutionary advancements in the health care industry which include Remote patient diagnosis, Remote Surgery and Distance Medical training. Remote surgery can be enabled by the use of medical robots with a main objective of treating rare medical cases, a few experienced doctors are able to perform remotely. The importance of such service will not only benefit hospitals in modern cities but also can be utilized to help people living in rural areas which expands the objective of Remote surgery knowing that hospitals in poor villages or cities are in fact suffering from the lack of medical expertise and even a remote consultation/diagnosis can be considered a great help. Another scenario might represent a health emergency where a patient in a critical condition need to be operated on immediately and cannot be transported to the hospital. Distance medical training can also be a usage scenario of uRLLC where actual operations can be taught by a mentor or a doctor specialized in a rare medical field to students located in another country.

Enabling such technologies would require a reliable network that is able to handle live video and audio streaming in addition to the haptic feedback triggered by the sensors installed on the patient to provide the remote surgent the required feeling of the medical robot/hand actions which plays an important role in decision making and precision.

The criticality of such application can be justified in that the control data that hold the instructions to the medical robot must be transmitted in ultra-reliable fashion because even a small amount of delay can lead to catastrophic effect on the patient’s health. The criticality of Remote Surgery is self-explanatory and thus, the required latency is 1 ms and the reliability level is 99.999% [10][11].

2.1.2 Industrial Automation

The second domain in which uRLLC is relied on is Control Systems. It is well-known that industries converted into automation for the sake of maximizing their productivity where humans are replaced to achieve accuracy, availability, and reliability in the manufacturing process. uRLLC usage scenario in this domain include the automation of product assembly lines, status reports of machineries, management of smart power grids in addition to process surveillance. Some of these applications require a latency of 0.5 ms and a reliability level with Block Error Rate (BLER) of about 10^{-9} which is usually achieved using wired networks [14]. Nevertheless, the deployment of wireless networks would provide more flexibility and reduce manufacturing and maintenance expenses which is a huge advantage compared to wired networks.

2.1.3 Intelligent Transportation Systems

uRLLC is expected to be a key enabler of some of the high technological transformations in transportation systems. One of them is Intelligent Transportation System (ITS) which includes several use cases such as Autonomous driving vehicles, Remote driving, Traffic management and Drone-based delivery. The main idea is fully connecting the vehicles in a way that enables the efficient communication and data exchange among them. This can be beneficial in terms of enabling these moving vehicles to interact with relatively complex situations with the help of each other in which the exchanged information can be relied on to provide extra realization to the traffic events happening in the area where the vehicle is moving at or heading to. This could empower road safety by preventing road accidents in addition to elevating the transportation efficiency where unwanted delays can be avoided once all vehicles can cooperate. As a result, a reliable communication service is required to enable the delivery of traffic events among the vehicles and the Roadside units (RSUs) with a low latency of 5-10 ms and a reliability level with BLER of 10^{-3} [10][11].

Table 2 lists some of the applications and their corresponding latency (ms) and reliability (%) requirements in addition to their estimated payload size (Bytes) and communication range (m).

Table 1. uRLLC Applications’ requirements [15][16][17][18][19]

Applications	**Reliability %**	**Latency (ms)**	**Payload Size (Bytes)**	**Com. Range (m)**
E-Health	99.999	30	28 - 14	300 – 500
Augmented Reality	99.999	0.4 - 2	12000 –	100 – 400
ITS	99.999	10 - 100	50 – 200	300 - 1000
Ind. Automation	99.9999999	0.25 – 10	10 - 300	50 - 100
Self-driving vehicles	99	1	144	400
Smart Grids	99.999	3 - 20	80 - 1000	10 - 1000
Process Automation	99.99	50 – 100	40 – 100	100 – 500
Tactile Interaction	99.99999	1	250	100000

2.2 Key Enablers of uRLLC

This section addresses some of the issues that complicate the implementation of uRLLC in addition to the technologies in which 5G NR is based on to tackle these issues.

Several research works focused on enabling uRLLC using different approaches. For instance, Authors in [32] addressed the theoretical principles in communication that supports the design of uRLLC. This included the Medium Access Control (MAC) protocols and interface diversity. Authors in [33] addressed issues related to the Physical layer and presented different technologies to tackle them including scheduling techniques and frame structure. The work in [34] highlighted the limitations in uRLLC and provided key directions to the next generation (Extreme uRLLC or xURLLC). The work included the prediction of traffic, Key Performance Indicators (KPI) and channel using machine learning approaches. It also included idea of joint communication and control design. This section includes an overview of the main issues tackled by 5G NR aiming to satisfy the above discussed strict requirements of uRLLC. These 5G features has been outlined by the 3GPP standard (Release 15).

2.2.1 Frequency Spectrum

The operation of 5G network over a range of 1-250 GHz of licensed and unlicensed spectrum is a key factor in enabling eMBB and uRLLC. Frequency bands between 600-700 MHz provide sufficient coverage for wide and indoor areas in addition to supporting high user mobility and higher user data rates than LTE cellular systems (30-250 Mbps) [20]. Frequency bands between 1 and 6 GHz provide a balanced performance between high data rate and wide coverage. Moreover, frequency bands between 25–39 GHz provide high data rates for specific applications that could reach up to tens of Giga bits per second (Gbps). Ultra-high frequencies or Millimeter bands also led to the deployment of new technologies including Massive Multiple Input Multiple Output (M-MIMO) antennas in addition to Ultra-Dense Networks (UDN) and Beamforming which are used to overcome the high pathloss of Millimeter Waves (mmWaves) (the radiation in Millimeter Bands) that affects the coverage [21].

2.2.2 Massive MIMO

5G is expected to use massive MIMO which is considered as a fundamental technology to support mmWaves. M-MIMO is essentially based on antenna array system that uses a large number of antennas on the transmitter and the receiver ends, providing better throughput and higher spectral efficiency. The number of antennas can reach up to 96 – 128 antennas. The main advantage of M-MIMO is to provide higher capacity in a wireless connection without the need to extra bandwidth. It enables the serving of large number of terminals using the same time and frequency resources [22].

2.2.3 Carrier Aggregation

Carrier aggregation was introduced by 3GPP specifications in Release 10. The technology is to allow the aggregation of up to 32 carriers (component carriers) which as a result increases the bandwidth significantly (3GPP Spec. Release 13) [23].

Carrier aggregation can be deployed in Time or Frequency Division Duplex. One of the disadvantages of this technology is the complex hardware circuitry in addition to the intermodulation product generated by receiving multiple signals in different frequencies which causes interference.

2.2.4 Dual Connectivity

Dual connectivity (DC) was designed to support heterogenous networks specifically, non-standalone 5G which is backward compatible with LTE. It provides the capability for the User Equipment (UE) of simultaneous transmission and reception of data on the same carrier component [24]. In order to achieve dual connectivity, the UE should be connected to more than one base stations (BSs) in which one is considered as a master BS and the other one as a slave BS. These BSs usually operate on different frequencies. The master BS is used to pass the control information while the user’s data is split among the master and the slave BSs. DC can provide higher throughput, robust mobility, and load balancing among the BSs. Dual connectivity can be implemented alongside carrier aggregation and is considered as one of the key solutions to support the reliability level required in uRLLC.

2.2.5 Channel Coding

Low density parity check (LDPC) channel coding scheme is used in 5G NR. LDPC plays a crucial role in providing a reliable connectivity. It is capable to provide low code rates below 1/3 which achieves high coding gains [25]. In addition, Polar code is also used in 5G for the purpose of control signaling in layers 1 and 2 [25]. It features low complexity decoding and can achieve Shannon capacity in various channel conditions.

2.2.6 Frame Structure

One of the unique features of 5G is the flexible frame structure it offers. The frame structure is a grid of time and frequency in which the frequency domain is divided into a number of Resource Blocks (RBs) depending on the available Bandwidth. Each resource block includes 12 subcarriers. Moreover, different numerologies are supported in 5G where each have different value of subcarrier spacing (SCS). SCS equals to 15 * 2^M KHz (M can take a value between 0 and 4) and ranges between 15 and 240 KHz. Higher Subcarrier spacing values are used for higher frequencies in order to reduce the Inter-Carrier Interference (ICI). The time domain consists of subframes where each subframe might contain one or more time slots. The duration of each subframe is 1 ms while the duration of the time slots is scalable as shown in Table 2. The scalability of the time slots duration is subject to the size of the subframe and size of the SCS where these slots must not cross the boundary of the subframe and can range from 0.125 to 1 ms. The time slot usually consists of 14 OFDM. Cyclic prefix (CP) is used with different lengths (depending on the SCS) in order to mitigate the effect of Inter-Symbol Interference (ISI). Another important feature is the scalable TTI as the number of OFDM symbols per TTI can vary according to the network preference. This feature enables the scheduling of UEs on Slot (14 OFDM symbols) and Mini-Slot (1-13 OFDM symbols) basis. TTI length can be adjusted by either reducing the number of OFDM symbols per TTI or by increasing the SCS and thus reducing the OFDM symbol duration. For instance, if the TTI is 0.125 ms, the UEs can be scheduled on slot based fashion with a SCS of 120 KHz or they can be scheduled in a Mini-Slot based fashion by using a SCS of 15 KHz and mini-slot size of 3 OFDM symbols. Mini-Slot based scheduling plays a crucial role in enabling uRLLC in 5G as the short TTI means a shorter processing time in addition to avoiding unnecessary delay, waiting to the next time slot for transmission. It is also worth mentioning that 5G frame structure can be aligned with the one in LTE by setting the subframes to be equal to 1 ms with 2 slots per subframe, 7 OFDM symbols per slot and a 15 KHz SCS. This allows backward compatibility for 5G and can enable the dual connectivity of both 5G and LTE. In 5G frame structure, each symbol can be used for either control, uplink or downlink transmission offering different time slot configurations as shown in Table 3 where "dl" means Downlink, "ul" means uplink and "f" means flexible (Downlink or Uplink).

Table 2. 5G NR Scalable Numerology [26]

Numerology	**0**	**1**	**2**	**3**	**4**
Subcarrier Spacing (KHz)	15	30	60	120	240
OFDM Symbol Duration (μs)	66.67	33.33	16.67	8.33	4.17
Cyclic Prefix Duration (μs)	4.69	2.34	1.17	0.57	0.29
OFDM Symbol including CP (μs)	71.35	35.68	17.84	8.92	4.46
Number of OFDM Symbols/Slot	7 or 14	7 or 14	7 or 14	7 or 14	14

Table 3. Examples of Slot formats for normal CP [27]

	Symbol Number													
Format	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
0	dl	dl	dl	dl	dl	d	dl	dl	d	d	dl	dl	dl	dl
1	ul	ul	ul	ul	ul	u	ul	ul	u	u	ul	ul	ul	ul
2	f	f	f	f	f	f	f	f	f	f	f	f	f	f
46	dl	dl	dl	dl	dl	d	dl	dl	d	f	f	f	f	ul
48	dl	f	ul	ul	ul	u	ul	dl	f	u	ul	ul	ul	ul

2.3 Overview of Heuristic Scheduling Algorithms

One of the key features of 5G is the Radio Resource Management (RRM) techniques that are utilized to improve system performance. One of the important parameters to achieve this desired improvement are Packet scheduling algorithms which play an important role in allocating resources represented by frequency and time to the connected users. These algorithms consider channel quality conditions and the Quality of Service requirements of these users when making the resource allocation decision aiming to provide an optimal tradeoff between system throughput, spectral efficiency, and fairness. These algorithms work at the base station and are responsible for allocating fractions of the spectrum to the connected users.

In this section, we provide a general overview of some of the schedulers used in our first proposed approach (Chapter 3) including their key features. These schedulers have different objectives and input parameters. We can categorize them into 1) channel unaware, 2) Channel aware/QoS unaware, and 3) Channel and QoS aware schedulers. To describe these schedulers, we use the notation $r_i(k,t)$ to describe the instantaneous throughput of user *i* at time *t* (The amount of data the gNB can transmit to user *i* at time *t* using RB *k*) as shown in equation (2.1) which is directly dependent on the channel condition of user *i.* Equation (2.1) expresses the well-known Shannon capacity which is the data rate in which the data can be transmitted in a reliable fashion (i.e. Small error probability).

$$r_i(t) = BW_i \log_2(1 + SNR_i) \qquad \text{Eq. 2.1}$$

Where BW is the available bandwidth and SNR is the Signal to Noise ratio.

$R_i(t)$ is the average throughput of user i during a fixed size time window tc [35] as shown in equation (2.2). The variation of tc from 0 to ∞ is used to control the utility function of the scheduler in which setting it to 0 would lead to maximizing the system throughput while losing fairness and setting it to ∞ would lead to maximizing the least average throughput and thus maximizing the fairness level among users [36].

$$R_i(t) = \left(1 - \frac{1}{tc}\right) * R_i(t-1) + \frac{1}{tc} * r_i(t) \qquad \text{Eq. 2.2}$$

2.3.1 Proportional Fairness in Time and Frequency (PFTF)

The main objective of this scheduler is to balance between maximizing the data rate and fairness among the users. It is considered as channel aware and QoS unaware scheduler. The utility function of this scheduler is written below[37][38]:

$$U = max_i \left[\frac{r_i(k,t)}{R_i(t) + \sum_{j=1}^{m} r_i(j,t)}\right] \qquad \text{Eq. 2.3}$$

Where $\sum_{j=1}^{m} r_i(j,t)$ is the total data rate of user i along all the resource blocks j to m which are allocated to it at TTI t, and $r_i(k,t)$ is the data rate of user i at resource block k.

2.3.2 Maximum-Largest Weighted Delay First (M-LWDF)

This scheduler is designed to serve real time users which unlike PFTF, it considers the QoS requirements of these users including their Delay budget D_{QoS} and the lifetime expiration of their packets (Packet Loss Rate) PLR_{QoS} [39]. Its utility function is written below:

$$U = max_i \left[Q_i * D_i(t-1)\frac{r_i(k,t)}{R_i(t)}\right] \quad \text{Eq. 2.4}$$

$$Q_i = \frac{-\log(D_{Qos})}{PLR_{Qos}} \quad \text{Eq. 2.5}$$

Where Q_i is the parameter that considers the QoS requirements of user i and parameter $D_i(t-1)$ is the Head of the Line packet delay addressed to user i.

2.3.3 Exponential/Proportional Fair (EXP/PF)

This Channel/QoS aware scheduler was designed for multimedia applications while prioritizing real time users over non-real time users. The utility function is written below[13]:

$$U = max_i \left[\exp\left(\frac{(Q_i * D_i(t-1) - QD}{1+\sqrt{QD}}\right) * \frac{r_i(k,t)}{R_i(t)}\right] \quad \text{Eq. 2.6}$$

$$QD = \frac{1}{N}\sum_{i=1}^{N}(Q_i * D_i(t-1) \quad \text{Eq. 2.7}$$

Other schedulers include Best Channel Quality Indicator (Best CQI) which aims to maximize the system throughout by selecting the UE with the best channel condition. Max-min allocates resources to the UE with the lowest data rate at time t. Round Robin provides the resources for each UE for a specific period of time and can achieve a perfect resource share fairness[40].

2.4 Related Work

In this section we introduce some of the related works that highlights the problem of resource allocation of eMBB and uRLLC traffic. Different techniques are studied in this section that involves both instant and reservation-based scheduling.

2.4.1 Optimal Instant scheduling

As mentioned in Chapter 1, instant scheduling aims to serve the incoming uRLLC traffic instantly by puncturing pre-allocated resources of other services (e.g. eMBB, mMTC). This approach has been addressed in the literature for both downlink and uplink transmissions.

In [41], the authors studied the downlink scheduling of eMBB and uRLLC traffic proposing a dynamic resource allocation scheme that deals with the stochastic nature of uRLLC traffic and aims to maximize the eMBB throughput while satisfying the requirements of uRLLC UE. The problem has been formulated as a convex optimization problem where the constraint is the uRLLC required probability of error forming a threshold that should not be exceeded. The authors examined different threshold values and studied the effect on the overall sum-throughput of eMBB UE. The stochastic traffic size of uRLLC has been transformed into a deterministic form using the Cumulative distribution function in which the traffic is generated using a pareto distribution. The main outcome of this work showed that the reliability level of uRLLC can impact eMBB UE significantly.

In [42], the authors propose an online joint scheduling framework algorithm of eMBB and uRLLC, formalizing and solving the problem of resource puncturing on eMBB traffic. The authors used different models to tackle this problem. The linear model is used when the degradation in eMBB data rates is directly proportional to the amount of punctured resources in which an optimal resource scheduling algorithm is introduced. The scheduler targets the stochastic nature of uRLLC traffic and aims to place it in a uniform random fashion in each slot while scheduling the eMBB UE via iterative greedy method that considers the expected degradation in eMBB data rates. The Convex model is used when the uRLLC traffic can be modeled as a convex function. The decomposition of this model is not as efficient as the linear model making an optimal allocation more difficult. This led the authors to adopt a simpler uRLLC traffic placement model which is fixed across the whole time slot (across all mini slots).

The authors characterized eMBB capacity regions and were able to derive the effective eMBB data rate after each puncturing process. They developed an approximated stochastic optimization algorithm for joint scheduling of eMBB and uRLLC traffic that aims to maximize the eMBB data rate. The third model is a threshold-based model that considers the data rate loss of eMBB users when a threshold is exceeded. This model allocates the resource uRLLC in proportion to either eMBB data rate loss or threshold loss aiming to minimize the probability of data rate degradation in an eMBB slot. The results show that the eMBB/uRLLC joint scheduling problem includes some features that enables an efficient decomposition of resources while satisfying service requirements.

In [43], the authors propose a downlink scheduling algorithm that aims to satisfy a minimum achievable eMBB data rate with an optimal allocation of resource for uRLLC traffic. The approach is based on resource puncturing that forces zero transmission power allocation for eMBB users by the base station. It addresses eMBB and uRLLC users with pending retransmissions with uRLLC having the highest priority in order to satisfy the reliability constraint. Maximizing the minimum eMBB data rate is based on two preferences in which the first one is the expected eMBB data rate till time slot t. The second preference is based on the uRLLC placement strategy which is derived according to historical uRLLC latency and reliability demands. The resource allocation decision is based on these two metrics and the results show a noticeable improvement of this approach over random resource allocation schemes.

In [44], the authors introduced a novel resource allocation algorithm that considers the control channel, latency, radio channel condition and the Hybrid Automatic Repeat Request (HARQ) in the user scheduling decision. The main idea is to avoid unwanted segmentation of uRLLC packets over several transmissions and to reduce the unnecessary queuing delay of uRLLC data as well. The criteria of this scheduling algorithm include the serving of pending HARQ retransmission ahead of buffered uRLLC data as they are already suffering from retransmission delay. The buffered uRLLC data are prioritized according to their latency requirements and how close they are to exceed it. Resource block where the uRLLC UE experience the best channel conditions are allocated first to provide more reliability and to reduce the possibility of more needed retransmissions. eMBB UE are then scheduled on the available remaining resource blocks using a proportional fairness scheduling algorithm. The results showed a 98% improvement in uRLLC latency and a 12% improvement in eMBB sum data rate, all compared to proportional fairness scheduling algorithm.

In [6], the authors used the same model used in [41] where they address the eMBB/uRLLC resource allocation as a probability constraint optimization problem that deals with the stochasticity of uRLLC traffic. In this approach, the eMBB UE are scheduled first according to a generalized proportional fairness (GPF) metric that involves their past data rates in addition to their instantaneous data rate. GPF is suitable for use when wanting to consider eMBB users with low channel quality as it takes the multiuser diversity when making the scheduling decision. The problem is modeled using a 2-Dimensions Hopfield Neural Networks (2D-HNN) including the investigation of the energy function. The results showed that the proposed approach can maximize the data rate of eMBB UEs while achieving the desired fairness level.

In [45], the authors investigate the use of unlicensed spectrum. The work includes a base station and a wireless access point where both have access to unlicensed spectrum when needed. The work investigate impact of using unlicensed spectrum by the base station on the wireless access point and propose a time-sharing based approach to tackle this issue. The problem of resource allocation is formulated as an optimization problem that aims to maximize the expected data rate of eMBB UE while satisfying the requirements of uRLLC traffic. The problem is solved using two heuristic scheduling algorithms due to its high complexity. The results indicate the achievement of a minimum rate of 15 Mbps compared to random RB allocation for uRLLC that achieve only 5 Mbps.

2.4.2 Optimal Reservation based scheduling

Reservation based approach as described in chapter 1 is based on pre-reservation of frequency channels for each type of service.

In [46], the authors addressed the resource allocation problem using Network slicing approach in which the problem was formulated as a risk-sensitive form which aims to enhance the reliability of eMBB and uRLLC traffic. A deep reinforcement learning approach has been adopted for maximizing the average data rate of eMBB UEs and minimization of eMBB data rate variance. The proposed optimization and learning scheme have multiple advantages in terms of solving the resource allocation problem. The results indicated that the proposed work could satisfy the requirements of uRLLC while preserving the desired reliability level of eMBB users.

In [47], the authors addressed the existence of eMBB and uRLLC in Cloud Radio Access Networks (CRAN) where network slices are given to both eMBB and uRLLC. A multicast and unicast transmissions are considered for eMBB and uRLLC respectively for the sake of improving the throughput of eMBB users. A revenue framework is proposed in which the operator’s revenue maximization problem is formulated as a mixed integer and nonlinear programming solved using convex approximation. The results show that the proposed work improves the system power consumption efficiency and revenue gain.

In [48], the authors address the Orthogonal Frequency Division Multiple Access (OFDMA) where the resources are virtually divided and distributed to eMBB and uRLLC services. The goal is to improve the spectral efficiency and to satisfy the reliability requirements of uRLLC. Thus, the problem is formulated as a Mixed integer programming in which the objective is to maximize the spectral efficiency while satisfying the requirements of both the eMBB slice and the uRLLC slice. An approximation of the problem was done to transform it into a convex optimization problem in which the objectives are combined using dual variables to form a Lagrangian function. Two resource allocation algorithms are proposed in which the first one is based on Powell–Hestenes–Rockafellar method and the second one is based on Branch and bound. A comparison with adaptive particle swarm optimization was used to evaluate the work and the results show improvement in the spectral efficiency and the reliability of uRLLC.

2.4.3 Sub-optimal scheduling schemes

In [55], the authors investigated possible solutions to the resource allocation problem among uRLLC traffic that involve the design of resource scheduler which benefits from low latency requirements of UE. The scheduler is able to increase the system capacity by almost 40% when compared to traditional schedulers like Exponential Proportional Fairness and Modified Largest Weighted Delay first. The authors tested several traffic models and evaluated their effect on fulfilling the uRLLC requirements when using the proposed scheduler.

In [49], the authors address the problem of reliability enhancement of uRLLC traffic. At first, they formulated the problem of resource allocation as an optimization problem in which it is non-linear, stochastic, and non-convex and the optimal solution would require a high computational complexity. As a result, a novel suboptimal scheduling algorithm with polynomial time complexity is proposed such that the main objective is to enhance the reliability level of uRLLC traffic. The method uses a sliding window model (equal to the slot length). The scheduling decision is based on the waiting time, the usability of the packets, the end to end delay budget, and the required reliability. The results indicate a remarkable improvement in the reliability level when compared to other methods (e.g. proportional fairness).

2.5 Main Contributions

The main contribution of this thesis is providing a low complexity resource allocation scheme that is based on resource puncturing. The approach adapts different utility functions by which fairness, maximum eMBB sum throughput and protecting users at the cell edge are the main objectives of the proposed approach. The utilization of heuristic scheduling algorithms is presented and evaluated to determine their usability in practice when dealing with uRLLC and eMBB at the same network. We also introduce an optimal allocation algorithm that focuses one preserving a certain fairness level among eMBB UE considering the amount of their punctured resources and their achieved data rates.

CHAPTER 3: SUB-OPTIMAL ALLOCATION OF URLLC/EMBB TRAFFIC IN 5G NETWORKS

Optimal allocation has been addressed intensively in the literature where the resource allocation problem is formulated and most of the times the solution to the problem is with high complexity. The uRLLC traffic as indicated in the previous chapters, has the stochastic nature and the strict requirements forces the BS to serve the traffic instantly. That is why low complexity solutions are considered more practical even with less efficiency when compared to the optimal approaches. At the BS, the resource allocation decision must be taken immediately, and the huge amount of calculations makes it impossible for the BS to cope with the traffic intensity and to satisfy the diverse requirements of different services.

In this section, we address eMBB aware scheduling methods for uRLLC with each having a different objective. All these methods are based on the resource puncturing approach in which the uRLLC is instantly served upon arrival. The main idea to test different methods for resource puncturing that would provide the best performance possible for both eMBB and uRLLC. The puncturing process is vital in determining the level of impact on every user and the decision of RB selection has a crucial role in elevating the efficiency of the puncturing algorithm.

Different parameters are considered in making the puncturing decision upon the arrival of uRLLC traffic. The channel conditions of both eMBB and uRLLC users represent the most important factor in the decision as it affects the users' data rates directly and knowing that each user might experience different channel quality at each RB. It is important to consider the state of the user at these RBs before puncturing in order to preserve fairness among eMBB users, provide better reliability for uRLLC and maximize the data rate of each eMBB user.

3.1 Best Resource Block for uRLLC.

The objective of the first method is to provide the best possible reliability level for uRLLC traffic considering the channel condition of the selected uRLLC UE. This is done by allocating Resource Block (RBs) with the best channel condition of the selected uRLLC UE. This method not only provides uRLLC UE with better reliability levels but also prevents the puncturing of extra resources in order to satisfy the latency requirements of the uRLLC traffic as better channel conditions mean a higher Modulation and Coding scheme value can be assigned to the uRLLC UE and thus more data can be transmitted using less resources. Slot boundary is taken into account and the method is updated once the RB is entirely consumed, moving to another RB here the uRLLC UE channel condition is the best compared to the other available RBs.

3.2 Protecting eMBB UE at the cell edge.

The second method aims to protect the eMBB UE at the edge level in order to prevent their starvation. Users at the cell edge most likely suffer from bad channel conditions and can never tolerate the effect of puncturing their resources. The CQI of each user is an important indicator that would help the BS to distinguish and apply protection policies that would lower the impact on these users. Protecting those users can be achieved by allowing the resource puncturing of eMBB UE with the best channel conditions as these users can be less affected by low uRLLC traffic density and their QoS level can be maintained even with the presence of uRLLC.

3.3 Maximization of eMBB Sum-Throughput

The third method aims to maximize the sum throughput of eMBB UE while satisfying the requirements of uRLLC UE. This can be achieved by targeting the resources of eMBB UE with lower channel conditions in order to protect the eMBB users with higher contribution to the overall sum throughput.

It can be noticed that the previous two methods might target the same eMBB UE in the case of having large uRLLC payload size or having multiple uRLLC transmissions at the same time slot.

For highest CQI and lowest CQI methods, the data rate of this eMBB UE at the punctured RB is updated according to equation (2.1). The channel quality is based on the SNR level of the UE over its assigned subcarrier. In this thesis, we used the approach proposed in [15] and [16] to calculate the CQI value of UE as a function of the SNR values of the selected user over all its assigned subcarriers. It is important to mention that the CQI reporting by UE is assumed to occur every 1 ms (1 Time Slot) which is vital for the algorithms to operate efficiently.

3.4 Knapsack inspired uRLLC fair punctured scheduling

The fourth method is a knapsack inspired scheduling method in which it aims to maximize the sum throughput of the eMBB UE while satisfying the requirements of the uRLLC traffic and preserving a fairness level among the eMBB UEs in terms of the amount of punctured resources from each of these users.

This method includes a number of objects representing the RBs in which each object has a profit and a weight associated with it. In each RB, the weight represents the data rate of the eMBB UE occupying it and the profit is the channel condition of the uRLLC UE at this RB. Better channel condition means that more data are being sent using this RB and thus more profit is gained.

The channel condition is measured using the Signal to Noise Ratio (SNR) of the selected uRLLC UE. The weight reflects the amount of impact on the sum throughput of eMBB UEs upon puncturing this RB. The knapsack is the constraint that needs to be considered when solving the problem and it represents the payload size of the uRLLC

UE that needs to be transmitted within the current time slot. The objective is to fill the knapsack in a way that maximizes the profit while considering the constraint.

The solution is given in a form of a set where each element shows if the RB has been selected for puncturing or not. The element value is between 0 and 1 which means that the RB can be partially punctured depending on the number of mini slots given to uRLLC UE.

In order to select the most suitable RB that gives us the best profit while considering its weight, we need to take the profit by weight ratio. To do that, channel conditions of uRLLC UEs and the achieved data rates of eMBB are rescaled in the range of 1 to 100 (this is to avoid any issues as the two parameters have different ranges). The RB with the highest profit by weight ratio is selected where a fraction of this RB is punctured and added to the solution set. The fraction of the RB represents the TTI duration of the uRLLC or the size of the mini slot (2,4,7 OFDM symbols). After each selection of RBs the payload size (knapsack) is updated based on the amount of data which we were able to transmit using this RB. This depends on the channel condition of the selected uRLLC UE at this RB. It can be noticed that one resource block can be entirely targeted throughout the whole process and starvation of certain eMBB UE is expected once their RB has a higher profit/weight ratio.

To prevent this and to provide a sense of fairness, we included a second constraint to the problem in which the resource block cannot be punctured two times in a row and the algorithm will move to another RB representing the second highest profit/weight ratio until the entire uRLLC payload is transmitted. The time complexity of the proposed knapsack problem is in the order $O(n \log n)$ which is acceptable in practical implementation. The methods' procedures are summarized in Figure 5.

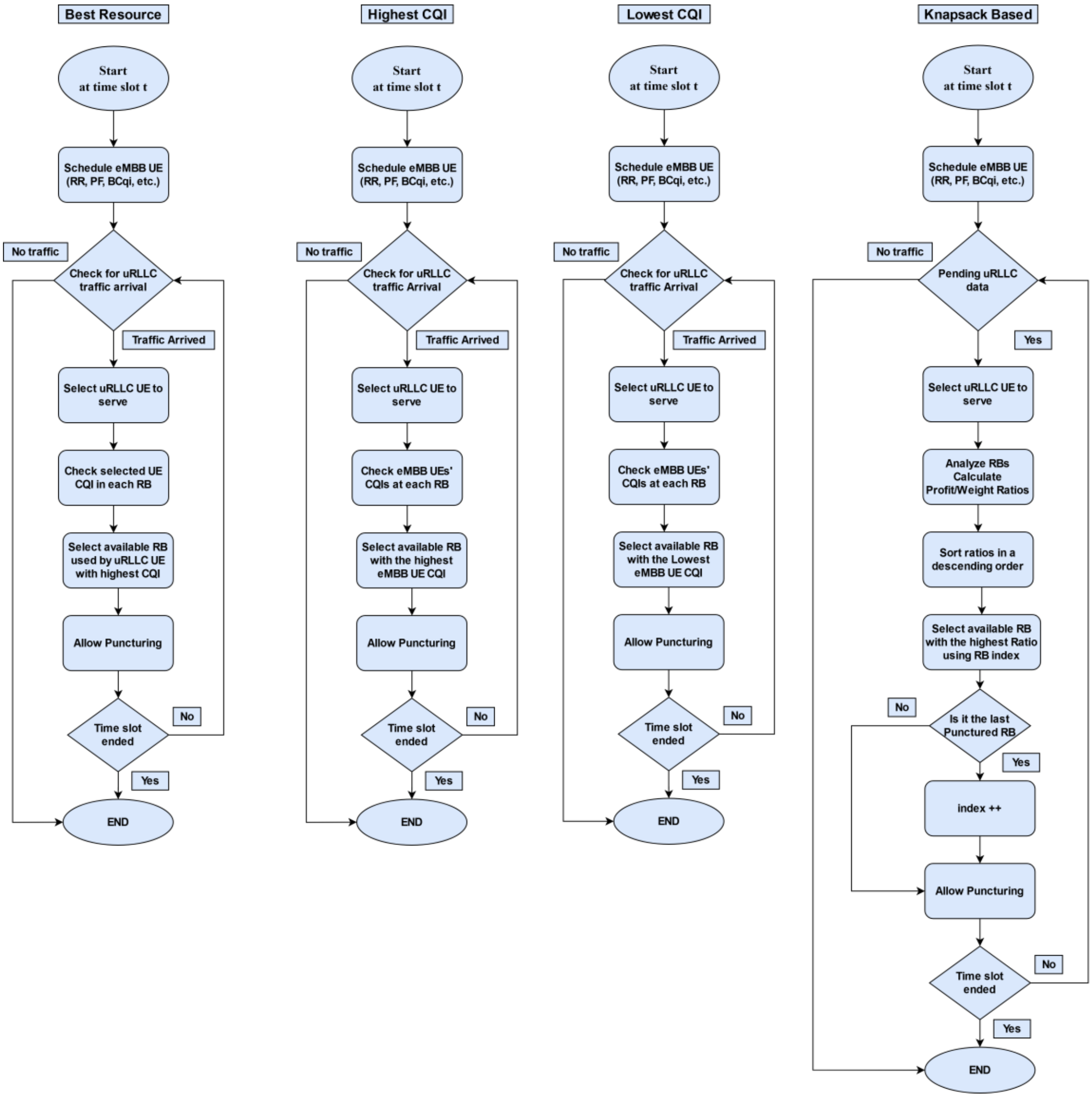

Figure 5: Proposed Punctured scheduling methods

These methods are designed to integrate heuristic scheduling algorithms discussed in Chapter 2 in order to make them suitable for use in the presence of uRLLC traffic. These methods are evaluated by measuring the performance of the heuristic scheduling algorithms when uRLLC traffic exists in different intensities. At time slot t, each method analyzes all RBs in order to find the most suitable one according to its criteria. All uRLLC UEs are considered to be using the same type of application where the latency requirement is 1 ms and all having the same priority. The knapsack inspired method is summarized as follows:

Table 4. Algorithm 1 Parameters

Parameter	**Meaning**
E	eMBB UE with allocated RBs at TTI t
U	uRLLC UE demanding immediate service.
SNR_U	Array of SNR values of uRLLC UE over all RBs.
SNR_E	Array of SNR values of eMBB UE over all RBs.
D_{size}	Payload size of uRLLC UE *s*.
R:	Array of eMBB UE instantaneous data rates at TTI *t*.
Ra	Array of eMBB Data rates and uRLLC SNR ratios.
P	Array used to keep record of the amount of punctured resources from each eMBB UE.
idx	Index of the last user with punctured resources.
N_{TTI}	Number of TTIs.
O_{RB}	Selected RB for puncturing

Algorithm 1: Knapsack inspired resource allocation scheme for eMBB/uRLLC traffic

```
Inputs: E, $U$, $SNR_U$, $SNR_E$, $D_i$
Output: $O_{RB}$.
idx=1;
for TTI=1 to $N_{TTI}$.
    while (any ($D_{size}$ > 0) && any ($P$<Number of Mini-Slots))
        Select a uRLLC UE ($U$).
        Calculate eMBB data rates at each RB using Eq. 2.1
        and store in $R$
        Calculate the ratio of $SNR_U$ and $R$ at each RB then
        store in $Ra$
        Sort $Ra$ in descending order
        if ($P(idx)$<Number of Mini-Slots)
            Puncture RB at $Ra(idx)$
            P=P+2
            $idx$=$idx$+1;
        end
        Update $D_{size}$ using equation 2.1 ($D_{size}$= $D_{size}$ - $r_u(k)$ ) where
        $r_U(k)$ is the data rate of uRLLC user $U$ at resource block $k$.
        if ($idx$>Number of RBs)
            $idx$=1;
        end
    end
end
```

CHAPTER 4: OPTIMAL ALLOCATION OF URLLC/EMBB TRAFFIC IN 5G NETWORKS

In this chapter, we address the resource allocation problem of uRLLC and eMBB while considering the optimal solution to the problem. The main aim as discussed in the previous chapters, is to protect eMBB UEs from the negative impact of uRLLC instant serving by the BS because of its high priority.

Optimization problems and the optimal solutions they provide represent the main design concept for wireless communication protocols especially in the resource allocation field. Resource allocation in the wireless communication domain faces many difficulties including the finite amount of resources and limited knowledge on the quality of these resources. Moreover, the demands of the different services discussed in chapter 1, makes it even harder to allocate the resources efficiently and to provide a dynamic response as these demands are also not deterministic to the BS. Also, adding them as part of the system design which is supposed to provide certain level of QoS guarantees to the users can be infeasible.

One important aspect of this chapter is to highlight the importance of considering the stochastic optimization and how the resource allocation problem is dealt with under uncertainty which is caused by the mobility of the users, continuously changing propagation environments in addition to the availability of the resources themselves.

The eMBB/uRLLC resource allocation problem in the literature is often formulated as an optimization problem that aims to maximize the data rate of the eMBB UE while fulfilling the requirements of uRLLC.

We use a similar problem formulation adapted from [41],[50],[51] with an added constraint that guarantees the desired fairness level among eMBB UE. Fairness

Index is used to control fairness level. The optimal allocation problem takes into consideration eMBB and uRLLC UE in which the optimization variables are the fractions of BW each service acquires from the BS.

The stochastic nature of the uRLLC can cause major issues to any optimization algorithm and in this type of formulation, uRLLC load is treated as a random variable derived from a random distribution (Poisson, Pareto, etc.) with different rate parameters which form a chance constraint optimization problem given the stochastic nature of uRLLC traffic.

A deterministic form of the uRLLC load can be calculated using the Cumulative distribution function (CDF) that corresponds to the random distribution used to represent the uRLLC flow. [41] [50] [51] considered the desired outage probability in which uRLLC must not exceed in their formulation.

Our contribution is to provide the desired fairness level by adding Jain's fairness index in the formulation which ensures a fair percentage of punctured resources from each eMBB UE. The fairness index forms a quadratic constraint that prevents us from transforming the problem into a convex form as the authors in the above references did. The problem formulation of our proposed work is shown below:

Maximize $\sum_{i=1}^{n}(R_i - R_{iu})\, log_2(1 + SNR_i)$ Eq. 4.1

Subject to $P\left[\sum_{j=1}^{m} R_{iu}\, log_2(1 + SNR_u) < L_u\right] \leq \gamma$ Eq. 4.1a

$\sum_{j=1}^{m} R_{iu} \leq BW$ Eq. 4.1b

$\frac{\left(\sum_{i=1}^{n}(R_i - R_{iu})\, log_2(1+SNR_i)\right)^2}{\sum_{i=1}^{n}[(R_i - R_{iu})\, log_2(1+SNR_i)^2]} > FI$ Eq. 4.1c

Where R_i and R_{iu} represent the amount of resources allocated and punctured to/from eMBB user i. SNR is the Signal to Noise ratio of eMBB user i. n is the number of eMBB UEs and m is the total number of mini-slots assigned to uRLLC UE. L_u is the size of uRLLC payload and γ represents the required reliability level of uRLLC traffic or it can be described as the confidence level for uRLLC users in which their data is transmitted within their latency budget (i.e. 1 ms) . BW is the available bandwidth.

In order to add the constraint that would ensure a level of fairness among eMBB UE, we need to consider the data rate of each eMBB by which Jain's fairness index formula is used as a quadratic constraint added to the formulation in Eq. 4.1. Assuming that the term FI (takes a value between 0 and 1) is the desired fairness level among all eMBB UEs, then the fairness constraint can be formulated as in Eq (4.1c). As a result of adding the fairness index, a non-linear constrained optimization problem is created which can be solved using "Fmincon" solver in MATLAB that supports such type of problems. Fmincon is based on non-linear programming and can be used to find the minimum of a scalar function with certain defined constraints starting at an initial point.

In this approach we are assuming that the bandwidth is fully used by the eMBB UE and a uRLLC traffic arrives at each time slot with different payload sizes. The puncturing of resources is based on the size of the mini slots or the TTI of the uRLLC while the initial allocation of the resources is slot based with a length of 1 ms.

We aim to observe the effect of forcing a certain fairness level on the performance of the scheduler in terms of the sum throughput of the eMBB UE. In addition, several error probability thresholds were assumed for the sake of comparing our work to [50]. Moreover, the random distribution is not used in the second part of the simulation as we aim to compare the optimal approach to the sub-optimal one presented in the previous chapter, instead, we used the same payload size and the number of uRLLC UE used in the sub-optimal approach.

It is important to mention that one of the limitations of the Fmincon solver is that it can converge into a local maximum instead of a global maximum. In addition, providing a reasonable fairness level among eMBB users cannot always be feasible because of the variation in these users' channel conditions. In other words, eMBB users with bad channel conditions might not be able to achieve a data rate close to those with better channel conditions and forcing a certain fairness level on the optimization algorithm would lead into a serious degradation in the sum throughput of the eMBB users in addition to the spectral efficiency. This is caused by forcing the optimizer to lower the data rates (i.e. provide less resources) to eMBB users with good channel condition in order to the satisfy the fairness threshold found in the constraint of Eq. 4.1c as it is the only way to narrow the gap between the eMBB users with different channel conditions in terms of achieved data rates.

This led us to test another idea which is mainly about providing a minimum data rate for each eMBB UEs. This idea can protect eMBB UE at the cell edge from starvation and provide them with acceptable data rates. Moreover, this method would also elevate the sum-throughput by not limiting the achievable data rates of eMBB UEs with good channel conditions.

This method includes the removal of Eq. 4.1c as a constraint and redefining Eq. 4.1a. The constraint $P\left[\sum_{j=1}^{m} R_{iu}\, log_2(1+SNR_u) < L_u\right] \leq \gamma$ can be transformed into a deterministic form using CDF which helps in avoiding the complexity that comes along with any stochastic variable. This method can be quite inefficient as the deterministic form can sometimes be very complex depending on the random distribution in which the random variable is derived from and the CDF of this distribution.

In our case, a Pareto distribution is used to produce the uRLLC load. This would enable us to work with a relatively simple CDF outcome that can be easily relaxed in our optimization process.

The idea is that if X is a pareto random variable, we can calculate the probability that X is greater than a value x. The CDF of the pareto distribution is given below.

$$F_X(x) = \begin{cases} 1 - \left(\frac{x_m}{x}\right)^{\alpha} & x \geq x_m \\ 0 & x < x_m \end{cases} \qquad \text{Eq. 4.2}$$

where x_m is the minimum positive value of x and represents the scale parameter of the pareto distribution. α is a positive value that represents the shape parameter of the pareto distribution. We can apply Eq. 4.2 on the constraint 4.1a as shown below: let us assume that the term u represents the outage probability of the uRLLC users.

$$u = P\left[\sum_{j=1}^{m} R_{iu}\, log_2(1 + SNR_u) < L_u\right] \leq \gamma \qquad \text{Eq. 4.3}$$

Then, we apply Eq. 4.2 on u as shown below.

$$P\,[u < L_u] \leq \gamma \Leftrightarrow 1 - F_X(u) \leq \gamma \qquad \text{Eq. 4.4}$$

$$\Leftrightarrow F_X(u) \geq (1 - \gamma) \qquad \text{Eq. 4.4a}$$

$$\Leftrightarrow u \geq F_X^{-1}(1 - \gamma) \qquad \text{Eq. 4.4b}$$

$$\Leftrightarrow u^{\alpha} \geq \frac{{x_m}^{\alpha}}{\gamma} \qquad \text{Eq. 4.4c}$$

Here, $F_X^{-1}(1 - \gamma)$ is the inverse CDF of uRLLC load which is evaluated using the reliability level defined earlier that simply ensures the delivery of the uRLLC load with its latency budget regardless of the payload size. Eq. 4.4c shows how the uRLLC random payload size is transformed into a deterministic form based on a predefined reliability level γ. As a result, the constraint in Eq. 4.1a can be redefined as follows:

$$\left(\sum_{j=1}^{m} R_{iu}\, log_2(1+SNR_u)\right)^{\alpha} \geq \left(\frac{x_m}{\gamma}\right)^{\alpha} \qquad \text{Eq. 4.5}$$

The formulation is now following a convex form and thus a global maximum can be achieved. The problem can be solved using CVX toolbox in MATLAB [54].

Maximize $\sum_{i=1}^{n}(R_i - R_{iu})\, log_2(1+SNR_i)$ Eq. 4.6

Subject to $\left(\sum_{j=1}^{m} R_{iu}\, log_2(1+SNR_u)\right)^{\alpha} \geq \frac{{x_m}^{\alpha}}{\gamma}$ Eq. 4.6a

$\sum_{j=1}^{m} R_{iu} \leq BW$ Eq. 4.6b

$(R_i - R_{iu})\, log_2(1+SNR_i) \geq \boldsymbol{r} \;\; \forall \mathrm{i} \in E$ Eq. 4.6c

Where E represents the eMBB users with RBs allocated to them at time slot *t* and r is the minimum data rate, prespecified for each of those eMBB UEs. Constraint in Eq. 4.6c would ensure a minimum data rate of $\boldsymbol{r}$ for each eMBB UE at time slot *t*.

CHAPTER 5: RESULTS AND DISCUSSIONS

In this chapter, all the results of our simulations are presented in addition to a detailed analysis on the performance of our proposed algorithms.

5.1 Simulation Settings and Evaluation Metrics

To evaluate the performance of the proposed algorithms, we induced a set of simulations that involved a varying number of eMBB UE randomly distributed around the base station and scheduled according to several schedulers defined in Chapter 2. It is assumed that all the resources have been previously allocated to eMBB UE. All simulations included a minimum uRLLC load size of 1 Mbps assuming a minimum uRLLC packet size of 32 bytes[51].

The uRLLC traffic is simulated as a Poisson flow with a rate parameter of 500. eMBB traffic represents a real time video streaming with an average delay threshold of 100 ms and a packet loss ratio threshold of 10 %. Resource puncturing is restricted to slot boundary and cannot exceed to the following slot.

The proposed algorithms are applied in each TTI and evaluated based on the performance of the heuristic scheduling algorithms in terms of average Throughput, Fairness and Spectral Efficiency. The evaluation metrics can be calculated using the below equations. Note that the data rate can be calculated using equation (2.1).

The first evaluation metric is the fairness of the scheduler which can be measured using the well-known, Jain’s fairness index [52] which can be used to determine if each user is receiving an equal share of resources compared to others. It can be calculated using the below formula which is a more generalized form of equation (4.1c) after removing the uRLLC punctured data and reorganizing the terms for more clarity.

$$Fairness\ Index = \frac{[r_i]^2}{N\sum_{i=1}^{N} r_i^2} \qquad \text{Eq. 4.7}$$

Where N is the number of users and r_i is the Data rate of user i.

In our simulations, we are adopting a full buffer model in which the eMBB users will always have data to transmit resulting in a full resource usage. Nevertheless, the amount of resources allocated to each user will surely differ according to each scheduler resulting in a different size of bits transmitted because of their different objectives and the channel condition each user is experiencing. That is why the second metric is important to measure how efficiently the bandwidth is used. Spectral efficiency [53] is used to measure the data rate that can be transmitted in a specific bandwidth through a cellular network and can be calculated using the below formula.

$$Spectral\ Efficiency = \frac{Sum\ throughput}{Total\ BW} \qquad \text{Eq. 4.8}$$

Table 5. Simulation Parameters

Parameter	Value
BS Max Power / UE	21 dBm
Cell Radius	1 km
Total Bandwidth	10 MHz
MIMO	2x2
Propagation Model	-128.1 + 37.6*log10(d) d: UE Distance from BS In Km
UE Distribution	Randomly Distributed
UE Noise Figure	7 dB
Number of Slots	1000
Sub-Carrier Spacing	15 kHz
Time Slot Size/Duration	14 symbols/1 ms
Slot Format	0[20]
eMBB Traffic Model	Full buffer
CQI Reporting	Every 1 ms

5.1 Simulation Results and Performance Evaluation

5.1.1 Performance evaluation of Heuristic Scheduling Algorithms

Different scenarios have been implemented in which we studied the impact of different uRLLC traffic intensities on eMBB UE. The first part of the results (Figures 6,7 & 8) can be considered as performance analysis of these heuristic schedulers which enables us to observe the behavior of each one in normal conditions (No uRLLC traffic).

In Figure 6, it can be noticed that Best CQI scheduler provides the highest system throughput among all other schedulers as it tends to select the UE with the best channel conditions. Better channel condition enables the base station to assign a higher MCS to the UE which leads to a better transmission efficiency.

As shown in Figure 7, Best CQI also achieves the best spectral efficiency because of its UE selection criteria that results a larger number of transmitted bits over the network compared to other schedulers.

Nevertheless, as shown in Figure 8 Round Robin, Proportional Fairness in Time and Frequency achieve a higher level of fairness among the users. RR fairness comes from the fact that it provides an equal amount of resource share to the users while PFTF takes into account the user’s previously achieved data rates in a given time window (Ex: 500 TTIs) in addition to the data rates achieved from allocated RB at the current TTI. Although RR is known with its fairness in resource allocation, it should be noticed that PFTF provides a better average throughput and the fairness is measured in terms of achieved data rates among the users resulting with a higher minimum achieved data rate over all its users when compared to RR. This feature is important when observing PFTF performance impacted by uRLLC and how it can manage to preserve a better data rate per UE compared to more advanced schedulers like M-LWDF and Ex-PF.

Ex-PF and M-MLWDF that are designed to satisfy the QoS requirements of real time and non-real time UE in terms of delay threshold and acceptable PLR. Their performance is considered better than most schedulers due to the fact that the UE QoS requirements is restricted by its channel quality (i.e. UE with bad channel quality would have a lower QoS) which results with the selection of UE with higher CQI and thus higher MCS is assigned to them by the base station which would explain the high throughput and spectral efficiency. In addition, the fairness level of these schedulers is significantly affected with the user density which can be observed from Figure 8. Considering the QoS requirements of the users forces the scheduler to prioritize some of them, which degrades the fairness level of the scheduler.

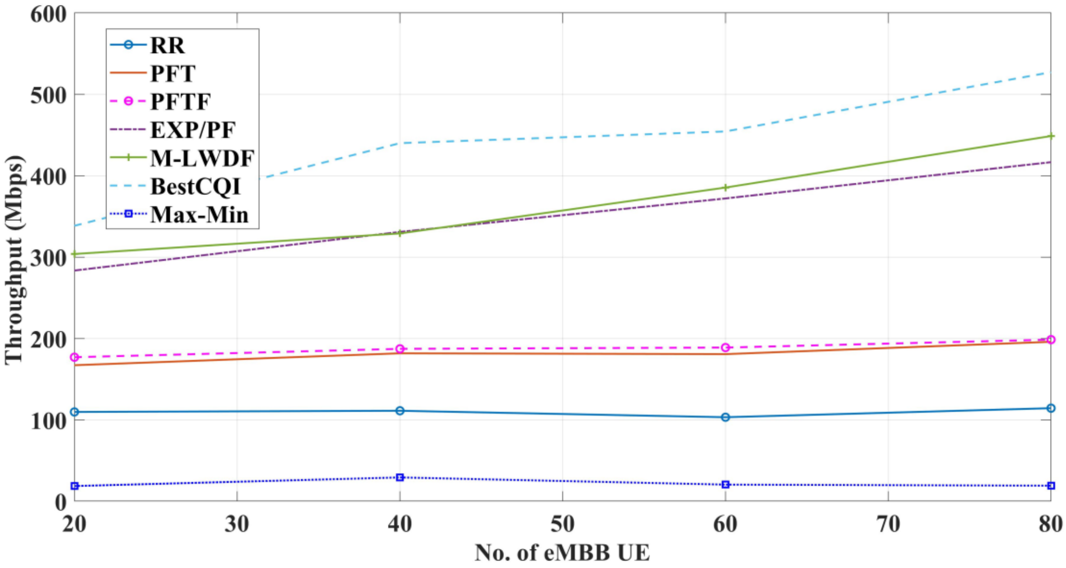

Figure 6. eMBB Sum Throughput Vs No. of eMBB UE

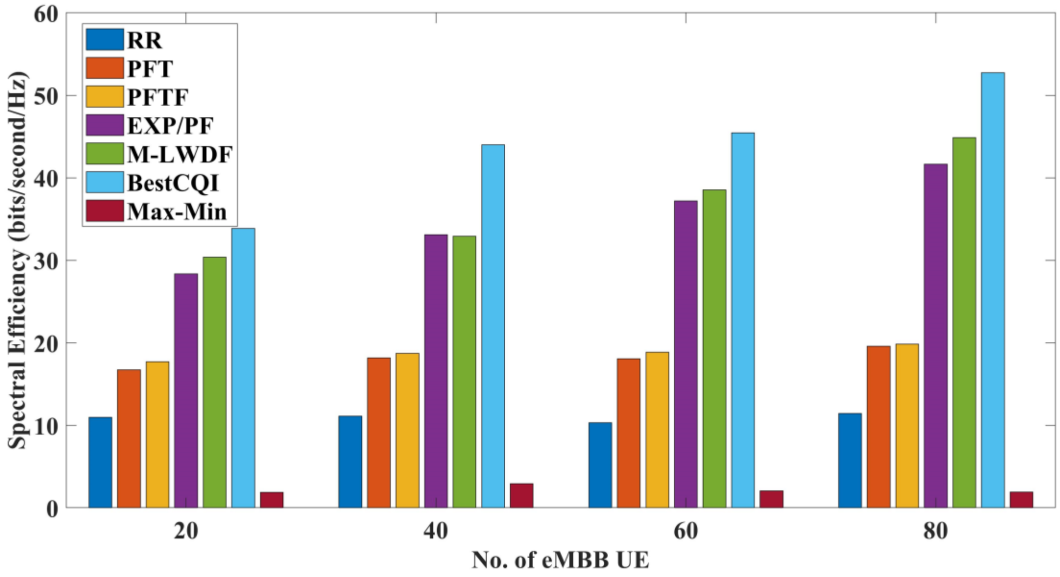

Figure 7. Spectral Efficiency Vs No. of eMBB UE

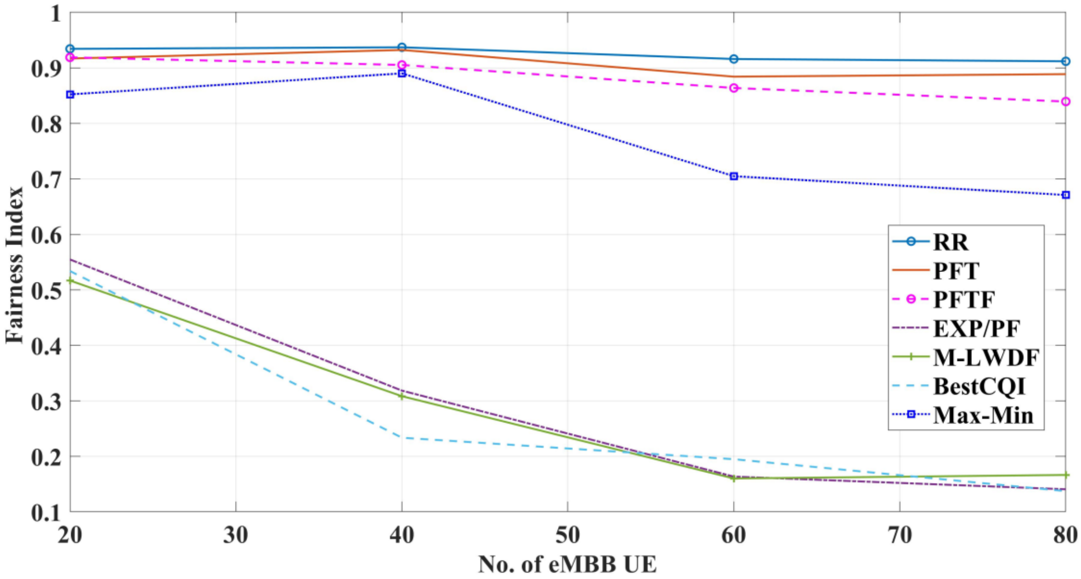

Figure 8. Fairness Index Vs No. of eMBB UE

The second part of the results discusses the performance of the proposed scheduling algorithms and evaluate the impact of uRLLC traffic on eMBB users. In order to observe this impact, we showed the performance of eMBB UE through time with no uRLLC traffic. This includes the fairness index, spectral efficiency, Sum throughput, and average throughput.

5.1.2 Best resource block puncturing algorithm

The first evaluation is done on Best resource block puncturing algorithm in which the RB where uRLLC UE experience the best channel condition is selected for puncturing with the main goal of providing high reliability level for uRLLC and to avoid the need of retransmission in case of lost packets. This method has a big advantage in which the better uRLLC channel condition would result in more data being sent at this resource block and thus, less resources are needed by the uRLLC traffic to satisfy its requirements in addition to lower impact of the performance of eMBB. Nevertheless, this approach does not consider fairness among eMBB UE and can be considered as eMBB unaware approach because it is possible that certain RBs will be consumed entirely by the uRLLC UE and certain eMBB UE will be impacted more as they occupy this RB. For comparison purposes, Figures 9, 10, and 11 show the performance of heuristic scheduling algorithms without uRLLC traffic during a specific time period.

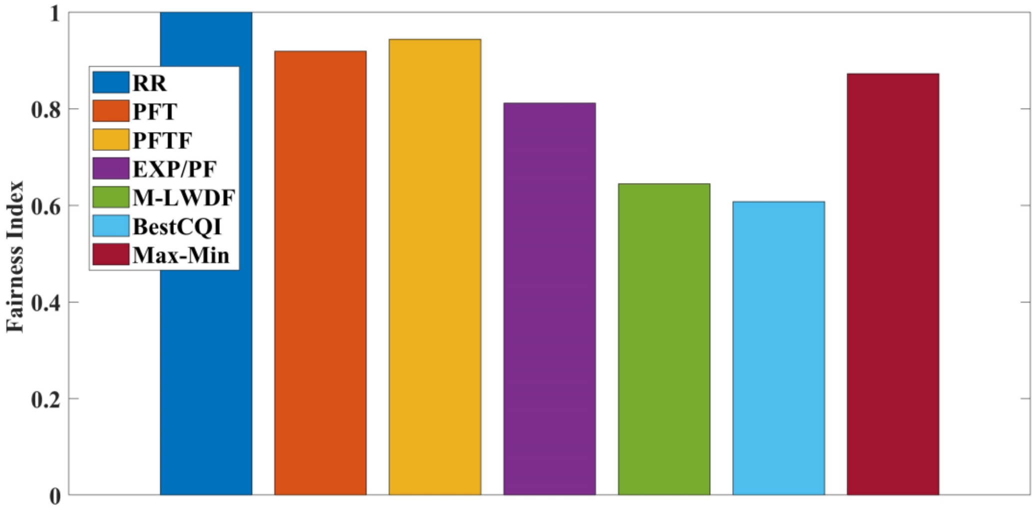

Figure 9. Fairness index (50 eMBB UE)

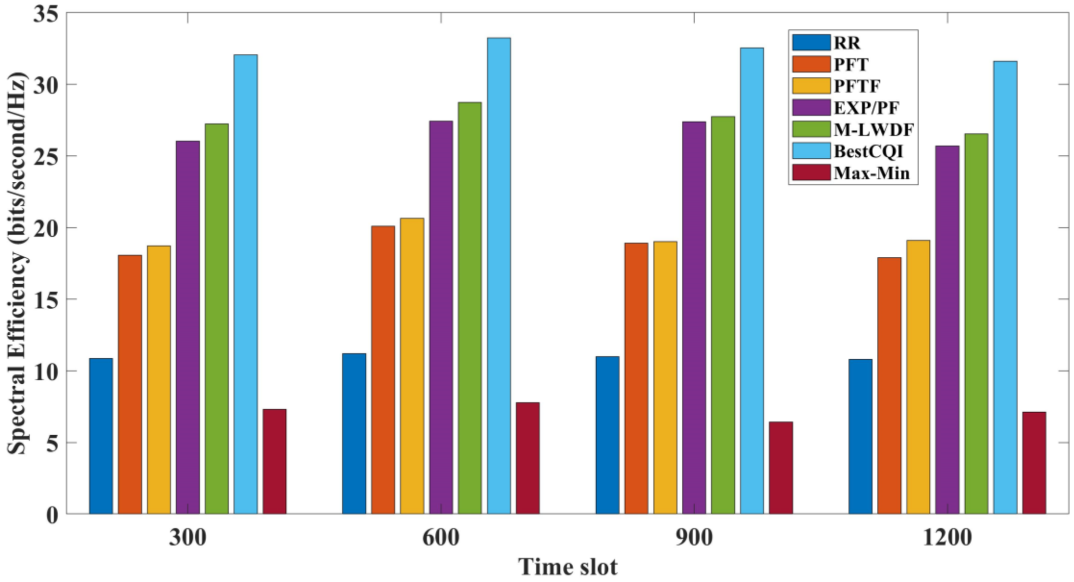

Figure 10. Spectral efficiency (50 eMBB UE)

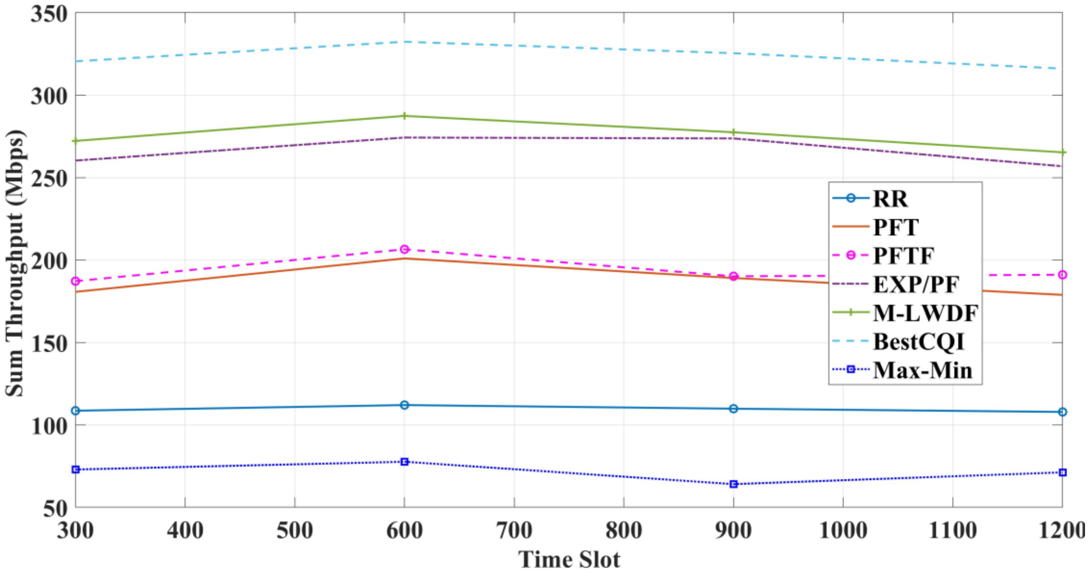

Figure 11. Sum throughput (50 eMBB UE)

The normal behavior of the heuristic scheduling algorithm is clear in these figures as explained earlier in this chapter. Figures 12 and 13 show the effect of uRLLC traffic arrival on these eMBB UE. The simulation of the two scenarios (with and without uRLLC traffic) has been performed at the same time which is important to make sure that the channel conditions are the same in both scenarios. The number of uRLLC UE increases after every simulation run.

Figure 12 shows the drop in spectral efficiency upon the arrival of uRLLC traffic. Our main metric when comparing the proposed algorithms is the level of degradation in each of the evaluation metrics when allowing uRLLC traffic with different densities to puncture the resources of eMBB UEs. Figure 12 shown a uniform degradation of spectral efficiency in the presence of 10, 20, 30 and 40 uRLLC UEs.

Figure 13, shows the sum rate when impacted by uRLLC and we can notice the huge drop in the performance of Best CQI, M-LWDF and Ex-PF as these algorithms consider the Quality of Service requirements of the served users which is (in the simulation setting) related to the channel quality of user over all of the user's assigned resource blocks.

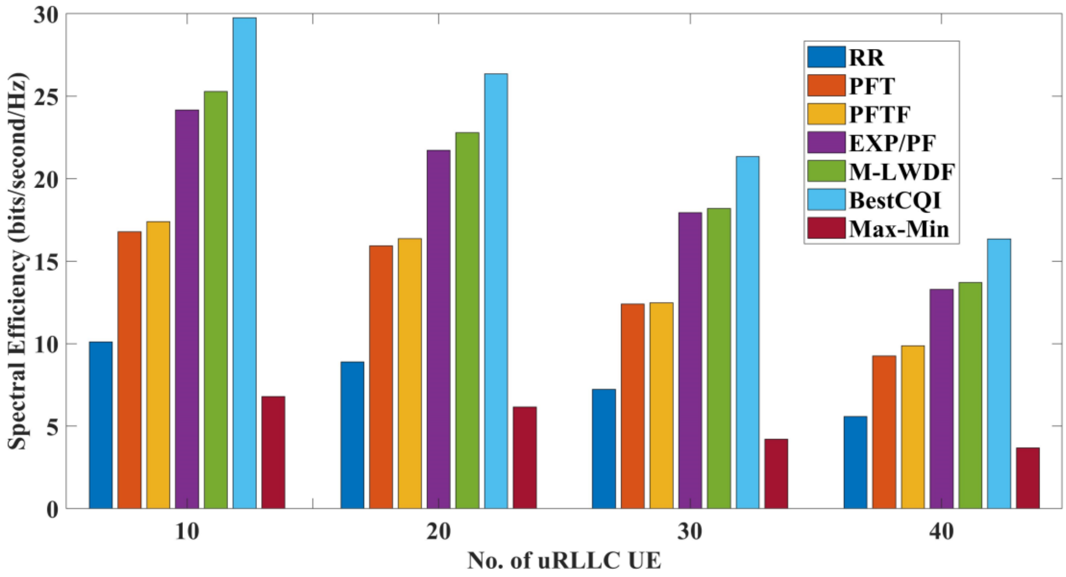

Figure 12. Spectral Efficiency Vs No. of uRLLC UE

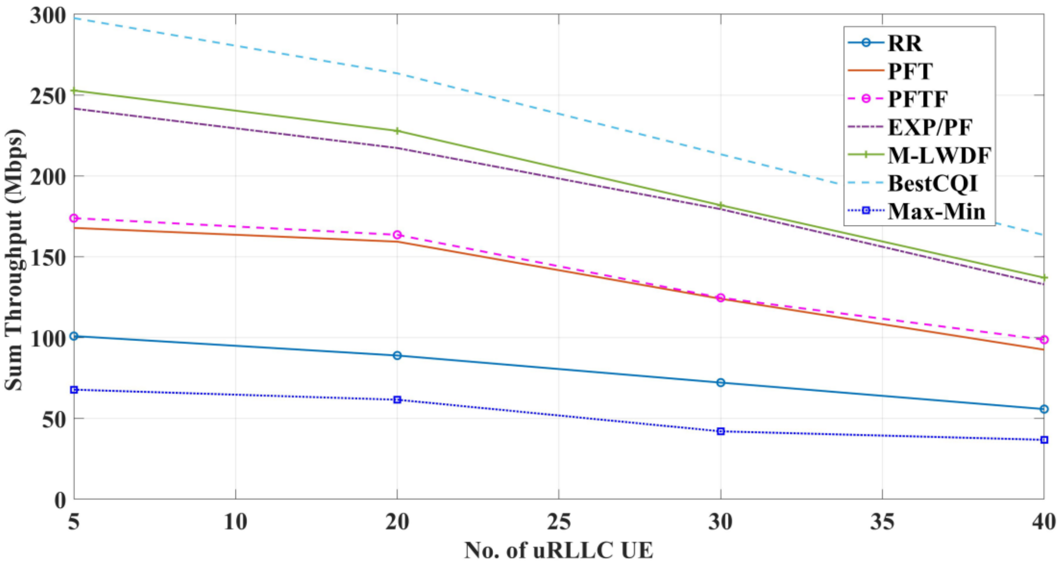

Figure 13. Sum Throughput Vs No. of uRLLC UE

5.1.3 Protecting eMBB UE at the cell edge

The second algorithm targets eMBB users with the highest achieved data rate across all RBs in time slot t. The main goal is to protect eMBB UE at the cell edge who most likely suffer from bad channel conditions and thus cannot tolerate the puncturing of their resources. Figure 14 and 15 show the performance of the proposed algorithms on the sum throughput of eMBB. The degradation is severe and the performance is much worse than Best resource puncturing algorithm discussed above as it tends to target eMBB UE with high data rate (i.e. best channel conditions) which impacts directly the sum throughput as these users have more contribution to the sum throughput more than others.

Figure 16 shows the effectiveness of this algorithm in protecting eMBB UE with bad channel conditions. As the figure shows, the percentage of punctured resources from those users are much less than those with better channel conditions. It is worth mentioning that the simulation included several runs by which the channel quality of eMBB users where fixed. This was done to observe the behavior of the heuristic algorithms and how efficiently can we integrate the proposed algorithm with these schedulers in order to deal with the co-existence of eMBB and uRLLC traffic.

The performance in terms of spectral efficiency as shown in Figure 17 is expected as the algorithm targets the UEs with higher data rate in which they have also more contribution in elevating the spectral efficiency when compared to others.

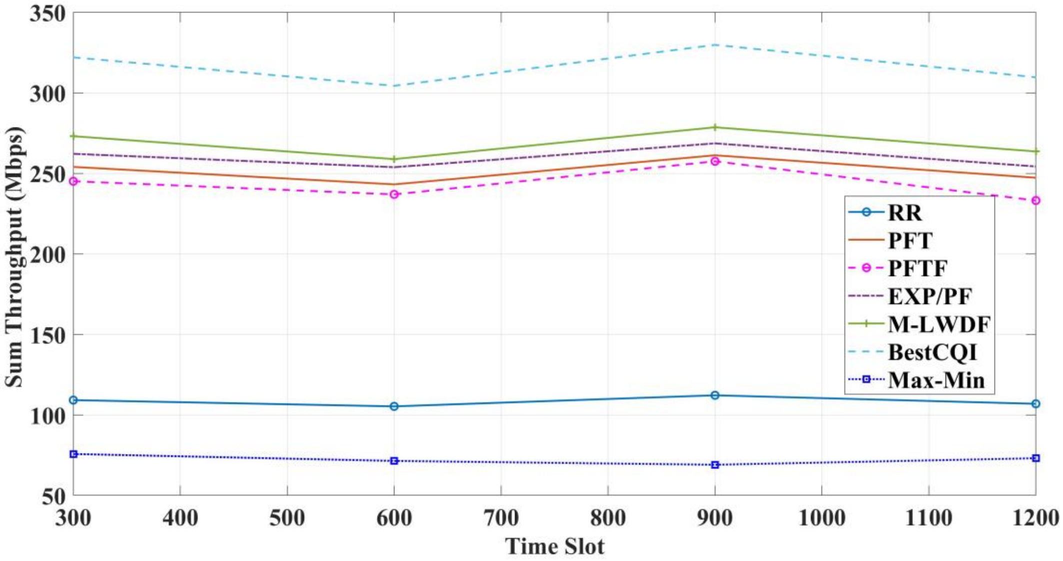

Figure 14. Sum Throughput (50 eMBB UE)

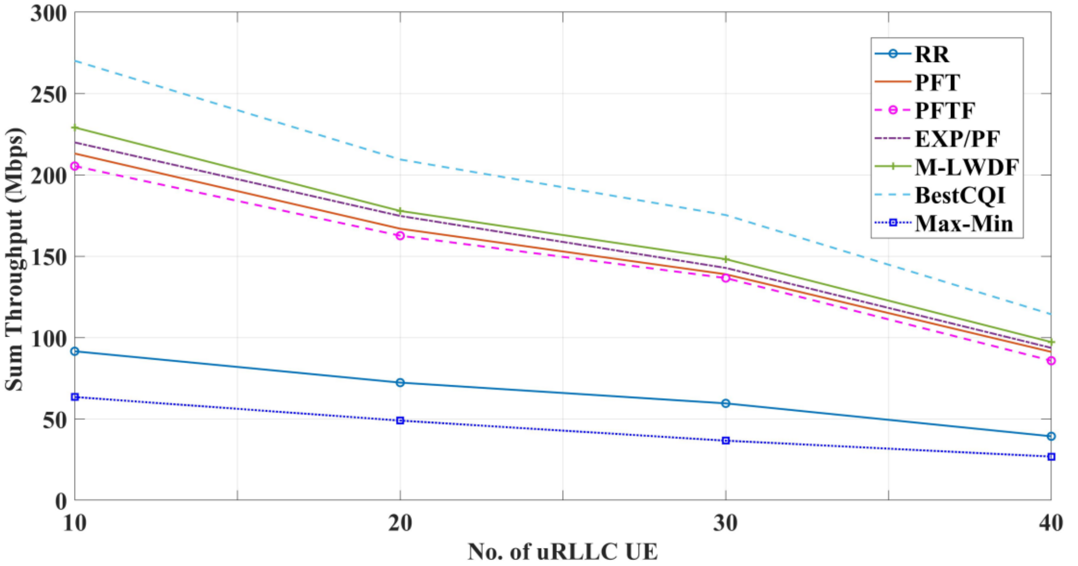

Figure 15. Sum Throughput Vs No. of uRLLC UE

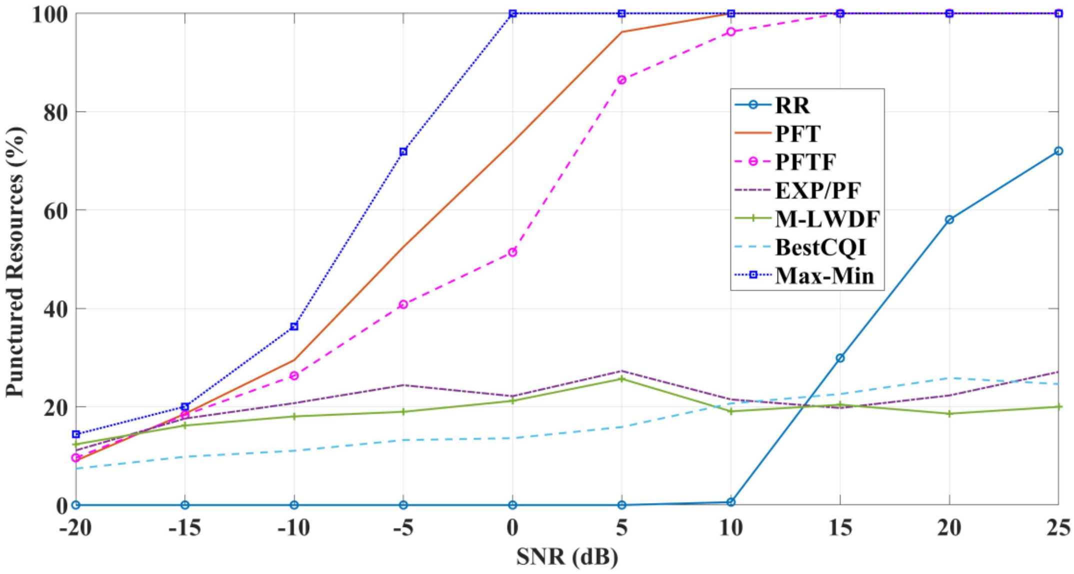

Figure 16. SNR Vs Percentage of Punctured Resources

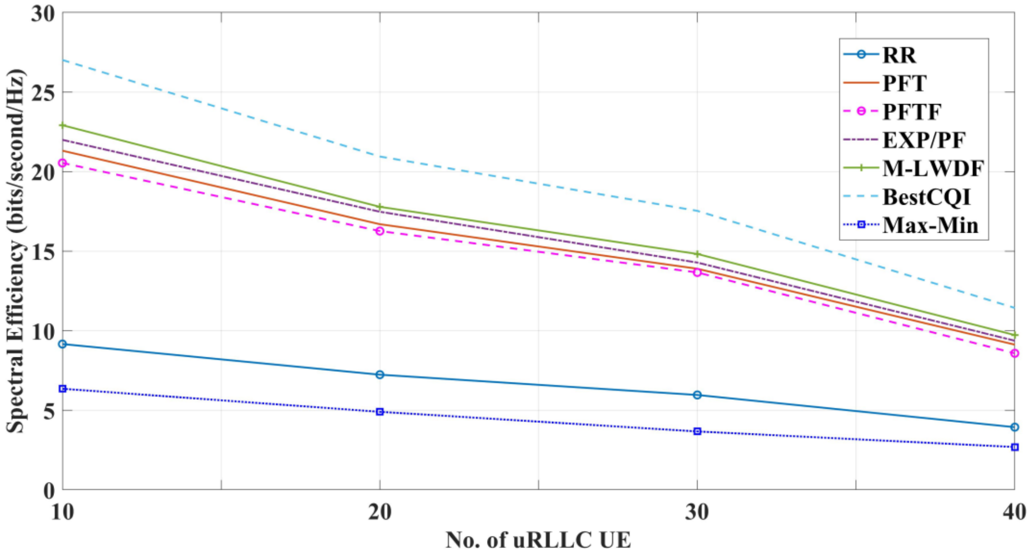

Figure 17. Spectral Efficiency Vs No. of uRLLC UE

5.1.4 Maximizing eMBB Data rate

The third algorithm aims to maximize the data rate of eMBB UE without considering the reliability level given to uRLLC and the fairness among eMBB UEs. The behavior of this algorithm tends to target eMBB UE with low channel conditions (i.e. have low contribution to the overall sum throughput) aiming to protect eMBB UE with high data rates. This would not only raise the sum throughput but also the spectral efficiency. Nevertheless, the algorithm does not take into account the possibility that these uRLLC UE might be experiencing bad channel conditions on the allocated RB and thus the need for more resources increases in order to satisfy the latency requirements of uRLLC traffic.

Figure 18 shows the performance of 50 eMBB UE with no uRLLC traffic and we can notice in Figure 19 that the degradation is less when compared to the previous two algorithms which is expected given the main objectives of the algorithm. Nevertheless, the efficiency of this algorithm can be considered acceptable when the evaluation is on the system level and the aim is to improve the overall performance without considering the state of the users and how they are affected. Figures 20 and 21 show that this algorithm can achieve a better spectral efficiency when compared to the previous algorithms in terms of degradation intensity.

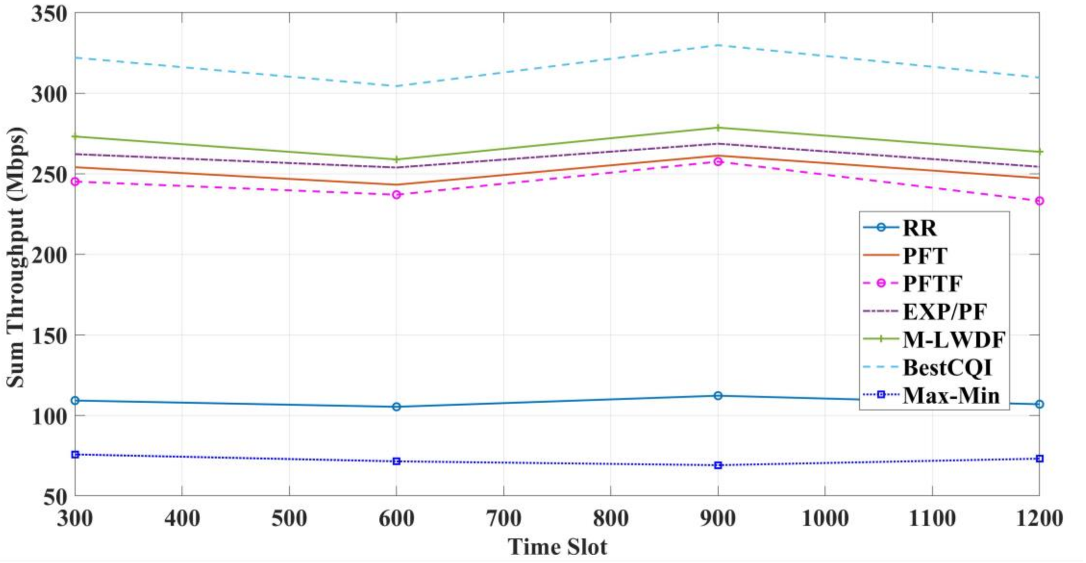

Figure 18. Sum-Throughput (50 eMBB UE)

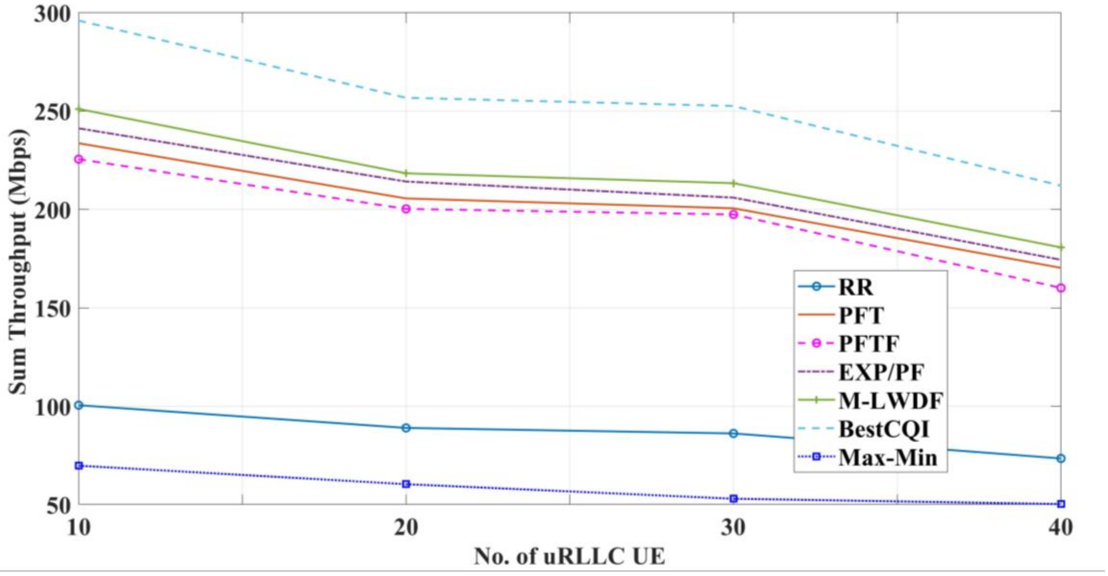

Figure 19. Sum throughput Vs No. uRLLC UE

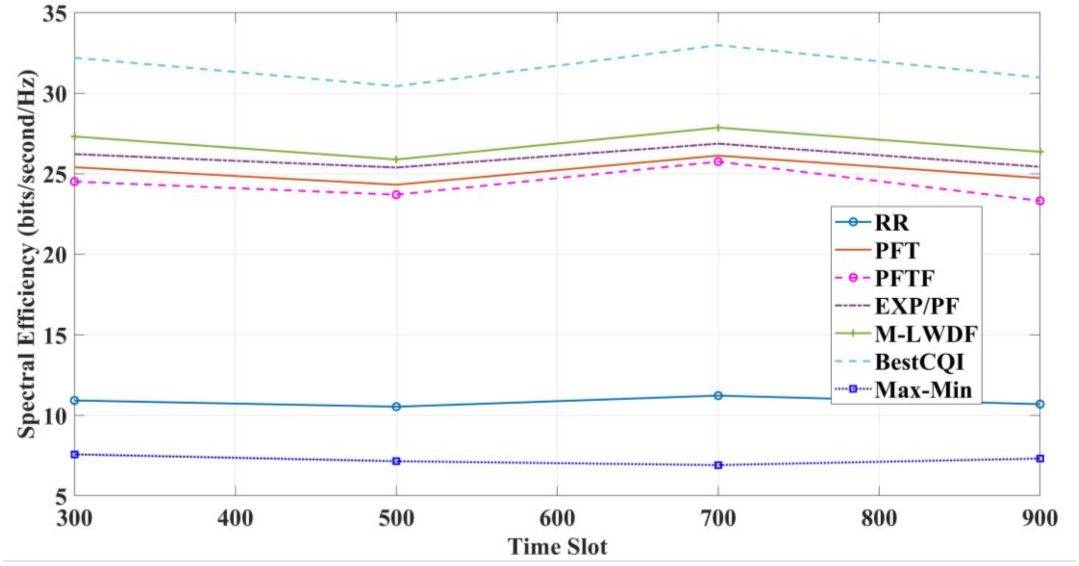

Figure 20. Spectral Efficiency (50 eMBB UE)

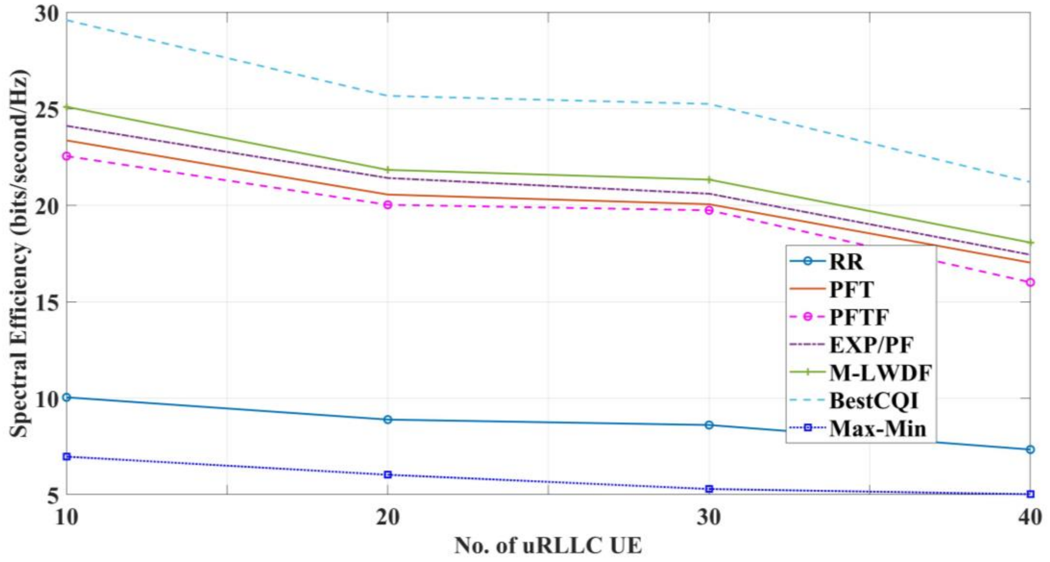

Figure 21. Spectral Efficiency Vs No. of uRLLC UE

5.1.5 Knapsack inspired puncturing resource allocation algorithm

This algorithm is discussed in detail in chapter 3. The main objective of this algorithm is to find the optimal resource block for puncturing by which we maintain an acceptable level of fairness in addition to achieving the best possible sum throughput of eMBB UEs. This algorithm considers the channel conditions of both eMBB and uRLLC UEs which gives it the privilege over the previously discussed algorithms. It simply tackles the limitations found in our results discussion above, trying to balance the tradeoffs between fairness level and sum throughput.

Figures 22 and 23 show the significance of this algorithm as it achieves the lowest degradation level in eMBB sum throughput when punctured by an increasing number of uRLLC UEs when compared to the previous algorithms. This is done by utilizing the knowledge of channel conditions of eMBB and uRLLC UEs in a way that leads to the selection of the most suitable RB for puncturing that would benefit both eMBB (does not degrade the um throughput) and uRLLC (provides an acceptable level of reliability).

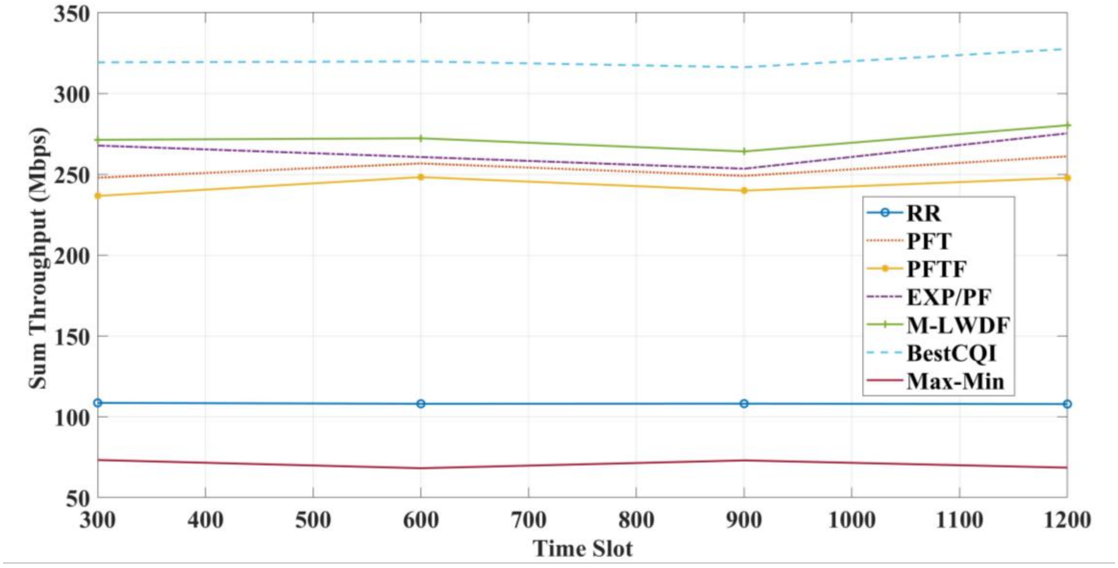

Figure 22. Sum Throughput (50 eMBB UE)

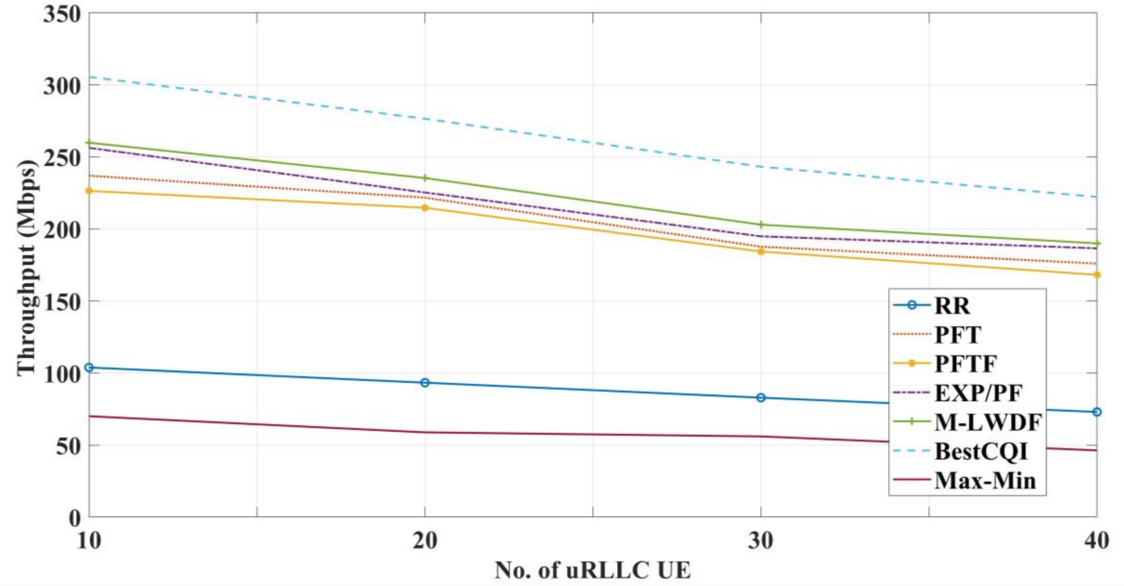

Figure 23. Sum Throughput Vs No. of uRLLC UE

Figures 24, and 25 show the degradation level on the spectral efficiency and it can be noticed that the degradation is not that significant given that the algorithm is capable of achieving high sum throughput of eMBB UEs.

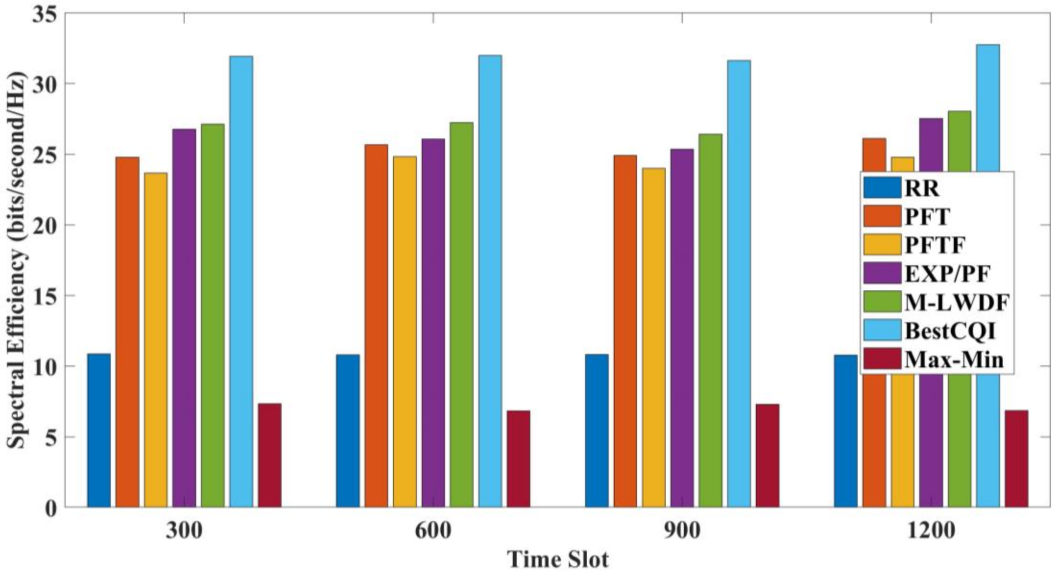

Figure 24. Spectral Efficiency (50 eMBB UE)

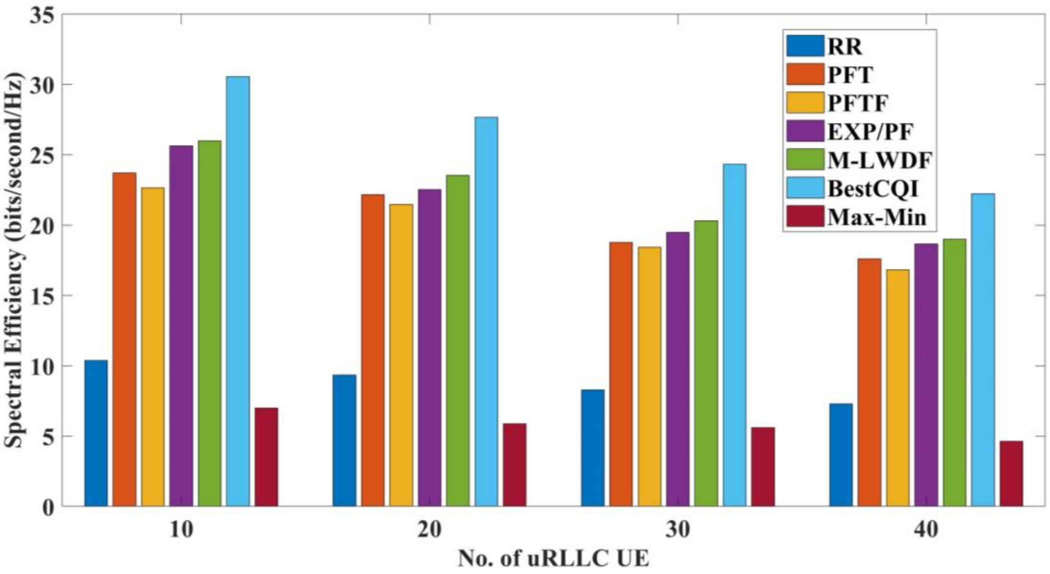

Figure 25. Spectral Efficiency Vs. No. of uRLLC UE.

Figures 26 and 27 show the fairness index of the heuristic scheduler. eMBB users with different numbers are required to measure the fairness of an algorithm in terms of the archived user data rates. Thus 20 to 80 eMBB users are used in this simulation. Figure 27 includes the puncturing of eMBB users by 40 uRLLC users with relatively small payload size and we can hardly notice the effect of this puncturing on the fairness level of all these schedulers. This indicates that the algorithm established a level of resilience against puncturing. This can be clarified by pointing out that the algorithm considers the achieved data rate of eMBB users in every RB and after each puncturing process which takes 2 OFDM symbols, this RB is also a potential target for puncturing in the next uRLLC allocation. Nevertheless, the algorithm forces the movement to another RB with the highest Profit/weight ratio to enforce a level of fairness in terms of the amount of punctured resources. Also, the data rate of the last punctured eMBB user is updated which lowers the profit in this RB and decreases the possibility of selecting it in future uRLLC allocations.

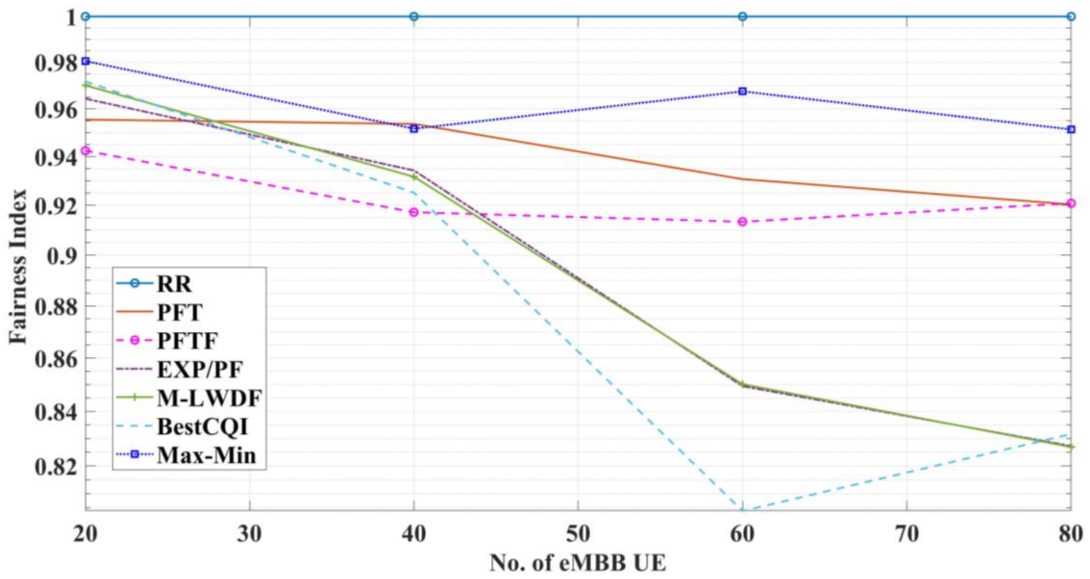

Figure 26. Fairness index (no uRLLC traffic)

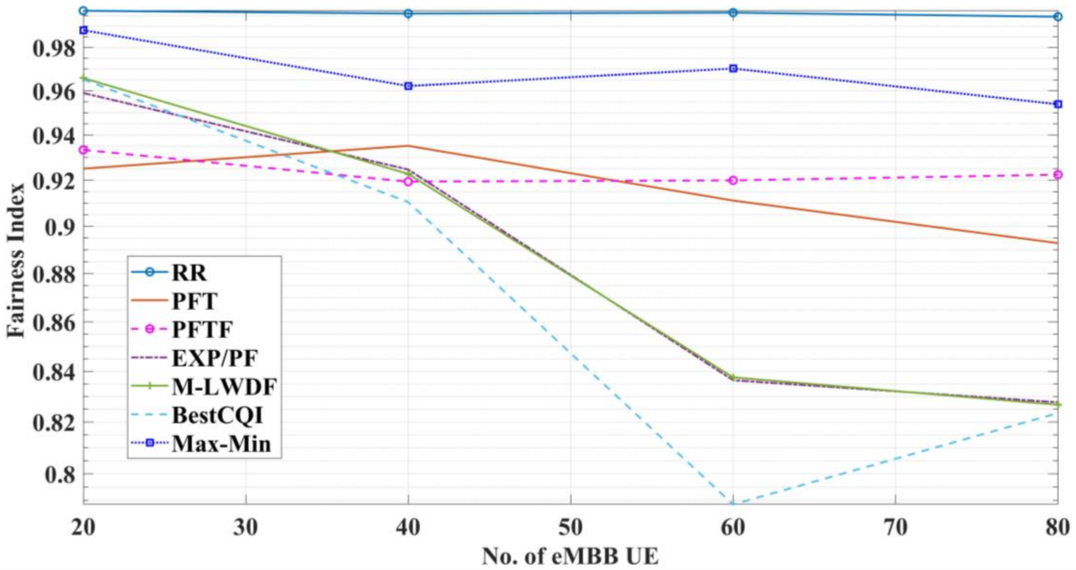

Figure 27. Fairness index (40 uRLLC UE)

Figure 28 shows the level of fairness even more clearly as it illustrates an almost fixed percentage of punctured resources among 10 eMBB UE. (the simulation included 10 eMBB UE and 20 uRLLC UE). This provide a stable service to these users which is an important factor in evaluating the pefromance of every algorithm.

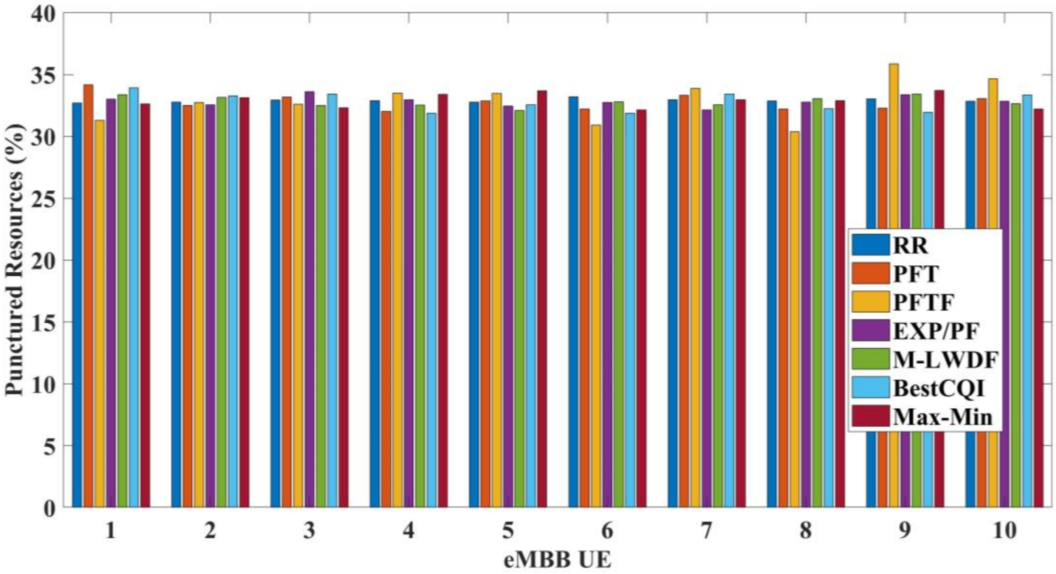

Figure 28. Percentage of Punctured Resources of 10 eMBB UE

5.1.6 Optimal Resource Allocation

The second part of this thesis includes the optimal allocation of resources conditioned by obtaining the desired level of fairness among eMBB UE. Figure 29 shows the impact of different fairness index values forced on the optimization problem and how it affects the sum throughput of the eMBB UE while being punctured by uRLLC traffic. The simulation included 20 eMBB and 20 uRLLC UEs in which performance is measured on slot basis.

It can be noticed that the more fairness level is required, the more impact is applied on the sum throughput of eMBB UE. This is because the optimization algorithm can no longer select eMBB UE with bad channel conditions (i.e. less contribution to the overall sum throughput) for puncturing and it is forced to treat the connected UE fairly according to the set fairness index value. The fairness index compares the achieved throughput of each eMBB UE with the other users in order to verify the fairness of the system by observing the difference between each user's data rate.

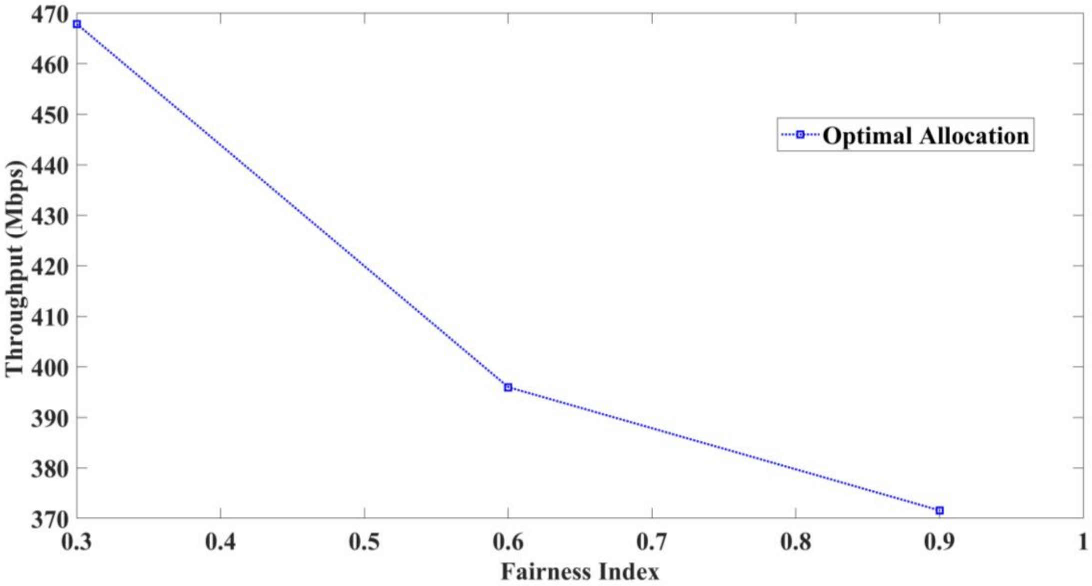

Figure 29. Fairness level Vs eMBB Sum throughput.

The second part of the simulation is to evaluate the effect of specified uRLLC outage probability on the eMBB sum throughput. It is important to mention that the simulation of the optimal allocation is done under the same simulation settings used in the sub-optimal approach except with different eMBB and uRLLC number of users.

Figure 30 shows how the sum rate declines with the decrease in the outage probability. A low outage probability provides more reliability level for uRLLC UE in a way that forces the optimization algorithm to assign better resources for those uRLLC users. In this part of the simulation, the fairness level is set to 0.5 which would allow us to observe the outage probability level effect on the sum rate with a medium level of fairness. It can be seen that the increase in the outage probability improves the performance in terms of eMBB sum-throughput.

It can be justified in the fact that the BS is no longer forced to allocate the best RB for the uRLLC in order to maintain the required reliability level which might create a conflict with the performance of eMBB users in the case where the RB of the uRLLC (with better channel conditions) is the same RB used by an eMBB UE with large contribution to the sum throughput.

Figure 31 shows how the proposed optimal resource allocation scheme is affected by different uRLLC traffic densities. The same number of uRLLC users are used in this simulation compared to Chapter 3. The degradation of the eMBB throughput is sharp but the overall achieved sum throughput is much larger than the sub-optimal approach.

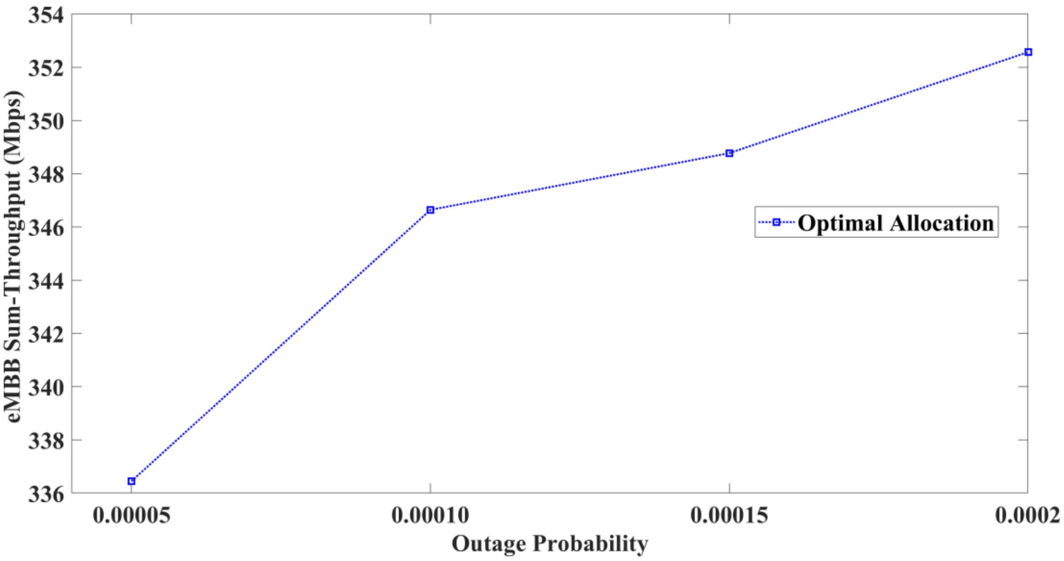

Figure 30. Outage Probability Vs Sum Throughput

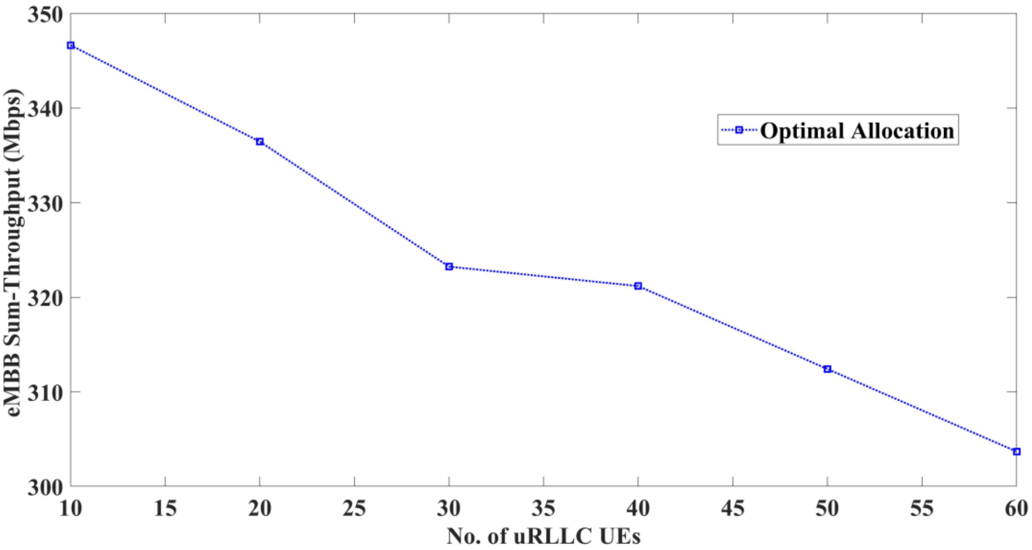

Figure 31. No. of uRLLC UEs Vs eMBB Sum Throughput

The below figures show the results of our second method that includes adding the condition of eMBB user's minimum data rate. Figure 32 shows the sum throughput of eMBB UEs when different outage probability values are required by uRLLC. The result is based on a minimum data rate of 60 Mbps for each eMBB UE and we can notice that the result is better than the result in Figure 30 where a fairness level among all users were required. This is due to the fact that the eMBB UE with bad channel condition might have a huge gap in terms of data rate when compared to eMBB users with good channel condition and adding a constraint that aims to narrow this gap will result in allocating resources to those with bad channel conditions in order to achieve that, resulting in a higher drop of eMBB sum-throughput. The results prove that a higher sum-throughput can be achieved once we allow the optimizer to allocate resources to eMBB UE with better channel condition and not required to close the gap between them and those with bad channel conditions.

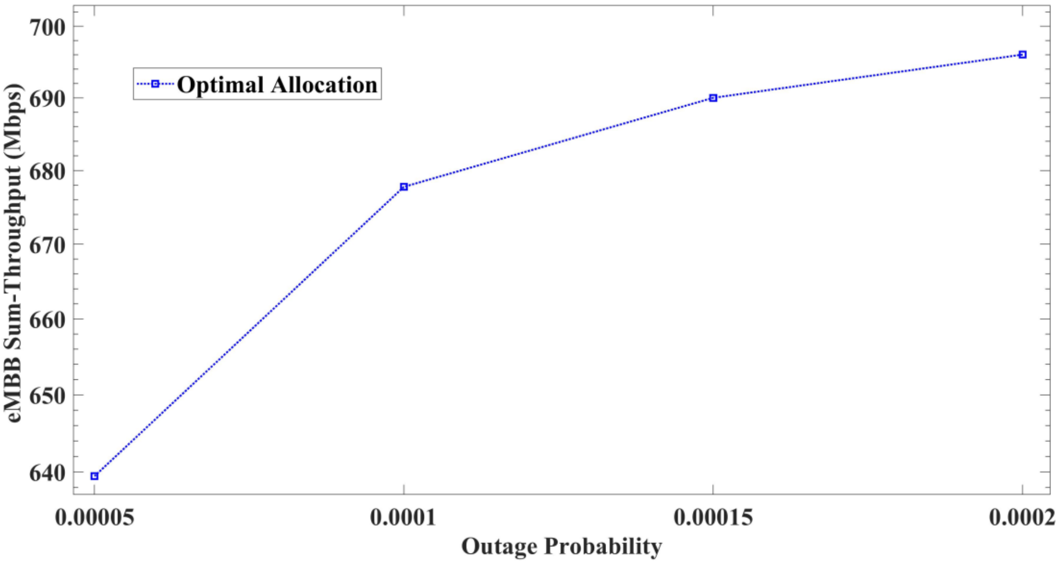

Figure 32. Outage Probability Vs eMBB Sum Throughput (with minimum data rate condition)

Figure 33. shows the impact of different number of uRLLC UEs on the sum-throughput of eMBB UE. The result includes the condition of maintaining a minimum data rate of 11 Mbps for each eMBB UE. This shows the advantage of adding this constraint instead if the fairness index as shown in Figure 29. The minimum data rate condition is more realistic and feasible in most scenarios that could include low, medium, or high uRLLC user densities. The degradation is acceptable given that the maximum achieved sum-throughput is higher than the one in Figure 29.

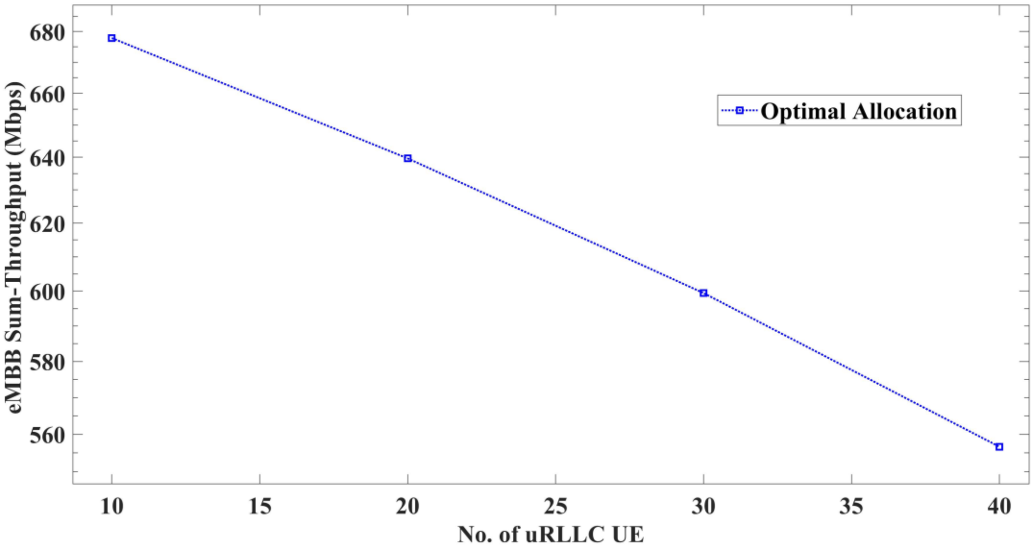

Figure 33. No. of uRLLC UE Vs eMBB Sum Throughput (with minimum data rate condition)

5.1.7 Sub-Optimal Vs Optimal resource allocation

In this section, we compare the four sub-optimal allocation algorithms to the optimal allocation algorithms in order to observe the gap between these two types of allocations. The suboptimal allocation included algorithm 1, which aims to provide the best resource block for uRLLC users, algorithm 2, which aims to protect eMBB users at the cell edge, Algorithm 3, which aims to maximize eMBB sum throughput without considering any other metrics and lastly, Algorithm 4 which is a knapsack inspired algorithm that aims to maximize eMBB sum throughput while maintaining a level of fairness. The Optimal allocation included two algorithms where the first aims to maintain a predetermined fairness level and the second one aims to guarantee a minimum data rate for eMBB users. Figure 34 shows a performance comparison among all proposed algorithms. It can be noticed that the knapsack inspired algorithm is in fact comparable to the optimal solution with a key feature of low complexity which demonstrates its effectiveness and feasibility for real life implementation. It outperforms all the proposed suboptimal allocation algorithms in terms of eMBB sum-throughput. Algorithm 1 provided a better performance with low user density compared to Algorithm 3. Algorithm 2 achieved the worse performance given that it targets eMBB users with the highest contribution to the sum throughput. The optimal allocation with guaranteed minimum throughput provided the best results compared to all others.

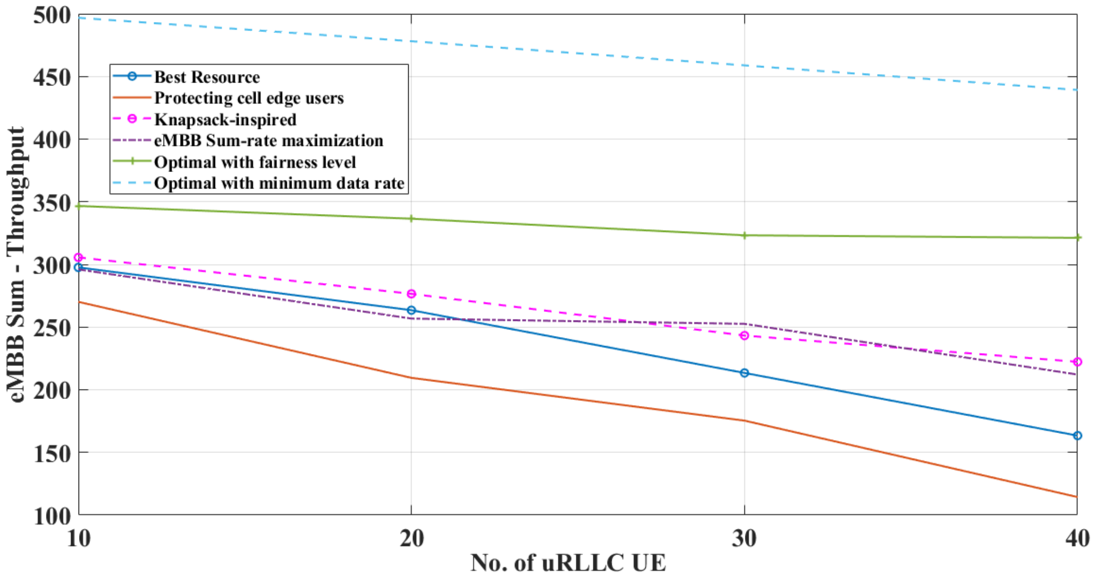

Figure 34. Sub-Optimal Vs Optimal Resource Allocation

CHAPTER 6. CONCLUSION

In this thesis we addressed the eMBB/uRLLC resource allocation problem in 5G NR. Two approaches have been proposed in which the first one includes a sub-optimal solution to the problem that features low complexity and an acceptable performance in terms of achieved eMBB sum throughput and fairness level. The approach consisted of a puncturing method that aims to use the knowledge of the users' channel conditions in order to make an optimal selection of RBs in the puncturing phase. The problem formulation is a knapsack inspired formulation in which the ratio of eMBB achieved data rates at each RB and the CQI of the uRLLC UE at each RB (e.g. at RB 1, we take the ratio of the data rate of eMBB UE occupying this RB and the CQI of the uRLLC UE at this RB) is used as a decision parameter to maximize the eMBB sum throughput while satisfying the requirement of uRLLC and providing the best possible reliability level at each time slot. A set of simulations have been conducted with uRLLC traffic of different intensities arriving stochastically and forcing the base station to puncture a previously allocated eMBB resources in order to achieve an instant serving of uRLLC traffic.

The proposed algorithm has been applied on top of heuristic practical scheduling algorithms and evaluated according to a set of performance metrics measured from these schedulers. Our first contribution is a performance analysis of the schedulers in order to study their behavior when dealing with different UE densities. The other sets of simulations included uRLLC traffics with a minimum intensity of 1 Mbps. The latency and the reliability requirements of uRLLC has been taken into account when served by the BS. Moreover, the transmission of each flow was based on the user's channel conditions and the desired BLER with no retransmissions allowed. We showed that the proposed algorithm can minimize the impact of uRLLC traffic on the schedulers' performance in terms of Sum throughput, Fairness, and Spectral efficiency. This work features a sub-optimal allocation that is comparable in terms of performance to the optimal allocation but with lower complexity which makes it suitable for real life implementation.

The second approach included an optimal resource allocation between eMBB and uRLLC services in addition to addressing the optimization under uncertainty. The formulation is based on transforming the stochastic uRLLC traffic into its deterministic form using CDF and aims to maximize the eMBB data rate while not exceeding a pre-determined outage probability threshold. The approach aims to also satisfy the desired fairness level among eMBB UE in terms of the percentage of punctured resources which is vital in protecting users with bad channel conditions as they are most likely considered a possible target in every optimal allocation method when maximizing the eMBB data rate is the objective. The results show the effectiveness of the proposed approach in maintaining the desired fairness level while achieving the better sum throughput compared to the sub-optimal allocation.

6.1 Limitations and future work

One of the limitations in our work is the fact that control signals are not considered in the simulations which might affect the overall performance of the proposed algorithms in terms of achieving a high sum-throughput and spectral efficiency. In addition, this work only focuses on the downlink transmission only of eMBB and uRLLC users which opens the possibility of future work that includes uplink transmission and how admission control of both eMBB and uRLLC UEs and how this could enhance the performance of the proposed algorithms by possibly not admitting eMBB UE who are not likely able to reach their QoS requirements in certain scenarios such as bad channel conditions, strict delay requirements or high uRLLC user density.

REFERENCES

[1] Ericsson AB, "Traffic Exploration Tool," Online tool. Available at: http://www.ericsson.com/TET/trafficView/loadBasicEditor.ericsson.

[2] M. Shahid, A. Morgado, K. Huq & J. Rodriguez. A survey of 5G technologies: Regulatory, standardization and industrial perspective, 2017. Digital Communications and Networks. 4. 10.1016/ j.dcan.2017.09.010.

[3] ITU-R, “IMT vision-framework and overall objectives of the future development of IMT for 2020 and beyond,” Recommendation M.2083-0, September 2015.

[4] A. Zaidi, F. Athley, J. Medbo, U. Gustavsson, G. Durisi, X. Chen, Chapter 1 - Introduction: 5G Radio Access, 5G Physical Layer, Academic Press, Pages 1-19, 2018, ISBN 9780128145784, https://doi.org/10.1016/B978-0-12-814578-4.00006-0.

[5] A. Osseiran, J.F Monserrat. & P. Marsch. 5G Mobile and Wireless Communication Technology. Cambridge University Press, 2016.

[6] 3GPP Technical Specification 38.824, “Study on physical layer enhancements for NR ultra-reliable and low latency case (URLLC),” Version 2.0.1, 2, March 2019.

[7] G. Pocovi, H. Shariatmadari, G. Berardinelli, K. Pedersen, J. Steiner, and Z. Li, “Achieving ultra-reliable low-latency communications: Challenges and envisioned system enhancements,” IEEE Network, vol. 32, no. 2, pp. 8–15, March 2018.

[8] 3GPP TR 38.913 V14.2.0, 5G; Study on Scenarios and Requirements for Next Generation Access Technologies (2017)

[9] M. Lema, A. Laya, T. Mahmoodi, M. Cuevas, J. Sachs, J. Markendahl, M. Dohler. Business Case and Technology Analysis for 5G Low Latency Applications. IEEE Access, 5, pp. 5917–5935, 2017.

[10] Ericsson. NR URLLC Rel-16 Use Cases and Requirements. Available online: https://list.etsi.org

[11] B. Holfeld, D. Wieruch, T. Wirth, L. Thiele, S.A. Ashraf, J. Huschke, I. Aktas, J. Ansari. Wireless Communication for Factory Automation: an opportunity for LTE and 5G systems. IEEE Commun. Mag, 54, 36–43,2016.

[12] 5G Americas. New services & applications with 5G ultra reliable low latency communications. [Online]. Available: https://www.5gamericas.org/wp content/uploads/2019/07/5G/5G_Americas_URLLLC_White_Paper_Final__updateJW.pdf

[13] H. Chen et al., "Ultra-Reliable Low Latency Cellular Networks: Use Cases, Challenges and Approaches," in IEEE Communications Magazine, vol. 56, no. 12, pp. 119-125, December 2018, doi: 10.1109/MCOM.2018.1701178.

[14] 5G for Mission Critical Communication–Achieve ultra-reliability and virtual zero latency, White Paper, Nokia Netw., 2016.

[15] G. J. Sutton et al., "Enabling Technologies for Ultra-Reliable and Low Latency Communications: From PHY and MAC Layer Perspectives," in IEEE Communications Surveys & Tutorials, vol. 21, no. 3, pp. 2488-2524, third quarter 2019, doi: 10.1109/COMST.2019.2897800.

[16] O. N. C. Yilmaz, “Ultra-reliable and low-latency 5G communication,” in Proc. Eur. Conf. Netw. Commun. (EuCNC), pp. 1–2, 2016,

[17] A.-M. Rahmani et al., “Smart e-health gateway: Bringing intelligence to Internet-of-Things based ubiquitous healthcare systems,” in Proc. IEEE Consum. Commun. Netw. Conf. (CCNC), pp. 9–12, Jan. 2015.

[18] O. Kocaba¸ S and T. Soyata, E-Health and Telemedicine: Concepts, Methodologies, Tools, and Applications. Hershey, PA, USA: Inf. Resources Manag. Assoc., Sep. 2015.

[19] M. Simsek, A. Aijaz, M. Dohler, J. Sachs, and G. Fettweis, “5G enabled tactile Internet,” IEEE J. Sel. Areas Commun., vol. 34, no. 3, pp. 460–473, Mar. 2016.

[20] J. Horwitz, "The definitive guide to 5G low, mid, and high band speeds". VentureBeat online magazine. December 10, 2019.

[21] A. Nordrum, K. Clark. "Everything you need to know about 5G". IEEE Spectrum magazine. Institute of Electrical and Electronic Engineers. January 27, 2017.

[22] E. G. Larsson, O. Edfors, F. Tufvesson and T. L. Marzetta, "Massive MIMO for next generation wireless systems," in IEEE Communications Magazine, vol. 52, no. 2, pp. 186-195, February 2014, doi: 10.1109/MCOM.2014.6736761.

[23] B. M. Hasiandra and Iskandar, "Planning and performance analysis of downlink inter-band carrier aggregation for LTE-Advanced 3GPP Released 13," 2016 10th International Conference on Telecommunication Systems Services and Applications (TSSA), Denpasar, pp. 1-6, 2016, doi: 10.1109/TSSA.2016.7871080.

[24] O. N. C. Yilmaz, O. Teyeb and A. Orsino, "Overview of LTE-NR Dual Connectivity," in IEEE Communications Magazine, vol. 57, no. 6, pp. 138-144, June 2019, doi: 10.1109/MCOM.2019.1800431.

[25] J. H. Bae, A. Abotabl, H.-P. Lin, K.-B. Song, and J. Lee, “An overview of channel coding for 5G NR cellular communications,” APSIPA Transactions on Signal and Information Processing, vol. 8, p. e17, 2019

[26] 3GPP Technical Specification 38.211, “5G; NR; Physical Channels and Modulation ,” version 16.2.0, Release 16, July 2020.

[27] 3GPP Technical Specification 38.213, “5G; NR; Physical layer procedures for control,” version 15.7.0, Release 15, October 2019.

[28] K. Pedersen, G. Pocovi, J. Steiner, A. Maeder. Agile 5G Scheduler for Improved E2E Performance and Flexibility for Different Network Implementations. IEEE Commun. Mag, 56, 210–217, 2018.

[29] G. Pocovi, K.I Pedersen, P. Mogensen. Joint Link Adaptation and Scheduling for 5G Ultra-Reliable Low-Latency Communications. IEEE Access, 6, 28912–28922, 2018.

[30] P. Popovski, K. F. Trillingsgaard, O. Simeone, and G. Durisi, “5g wireless network slicing for eMBB, URLLC, and mMTC: A communication theoretic view,” IEEE Access, vol. 6, pp. 55 765–55 779, 2018.

[31] 3GPP, “Technical specification group services and system aspects; release 15 description,” Tech. Rep., TR 21.915, v1.1.0, March 2019.

[32] P. Popovski, C. Stefanovic, J. J. Nielsen, E. de Carvalho, M. Angjelichinoski, K. F. Trillingsgaard, and A.-S. Bana, “Wireless access in ultrareliable low-latency communication (URLLC),” IEEE Transactions on Communications, vol. 67, no. 8, pp. 5783–5801, Aug 2019.

[33] H. Ji, S. Park, J. Yeo, Y. Kim, J. Lee, and B. Shim, “Ultra-reliable and low-latency communications in 5g downlink: Physical layer aspects,” IEEE Wireless Communications, vol. 25, no. 3, pp. 124–130, June 2018.

[34] J. Park, S. Samarakoon, H. Shiri, M. K. Abdel-Aziz, T. Nishio, A. Elgabli, and M. Bennis, “Extreme urllc: Vision, challenges, and key enablers,” arXiv preprint arXiv:2001.09683 (2020).

[35] S. Musleh, M. Ismail & R. Nordin. Effect of average-throughput window size on proportional fair scheduling for radio resources in LTE-A networks. 80. 179-183, 2015.

[36] W. Hsu. “Analysis of hybrid schedulers for CoMP resource allocation in LTE-Advanced SU-MIMO systems.” (2013).

[37] F. Afroz, K. Sandrasegaran and P. Ghosal, "Performance analysis of PF, M-LWDF and EXP/PF packet scheduling algorithms in 3GPP LTE downlink," 2014 Australasian Telecommunication Networks and Applications Conference (ATNAC), Southbank, VIC, pp. 87-92, 2014. doi: 10.1109/ATNAC.2014.7020879.

[38] E. Yaacoub; Z. Dawy, "Suboptimal Implementation with Proportional Fairness," in Resource Allocation in Uplink OFDMA Wireless Systems: Optimal Solutions and Practical Implementations , IEEE, pp.89-112, 2012. doi: 10.1002/9781118189627.ch7.

[39] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, and R. Vijayakumar, “Providing quality of service over a shared wireless link,” IEEE Communications Magazine, vol. 39, no. 2, pp. 150–154, Feb 2001.

[40] F. Capozzi, G. Piro, L. A. Grieco, G. Boggia, and P. Camarda, “Downlink Packet Scheduling in LTE Cellular Networks: Key Design Issues and a Survey,” IEEE Communications Surveys Tutorials, vol. 15, no. 2, pp. 678–700, July 2012.

[41] M. Alsenwi, & C. Hong. Resource Scheduling of URLLC/eMBB traffics in 5G New Radio: A Punctured Scheduling Approach, 2018.

[42] A. Anand, G. de Veciana, & S. Shakkottai. Joint Scheduling of URLLC and eMBB Traffic in 5G Wireless Networks. IEEE/ACM Transactions on Networking. PP. 1-14, 2020. 10.1109/TNET.2020.2968373.

[43] A. Pradhan and S. Das, "Joint Preference Metric for Efficient Resource Allocation in Co-Existence of eMBB and URLLC," 2020 International Conference on COMmunication Systems & NETworkS (COMSNETS), Bengaluru, India, pp. 897-899, 2020. doi: 10.1109/COMSNETS48256.2020.9027365

[44] A. Karimi, K. I. Pedersen, N. H. Mahmood, G. Pocovi and P. Mogensen, "Efficient Low Complexity Packet Scheduling Algorithm for Mixed URLLC and eMBB Traffic in 5G," 2019 IEEE 89th Vehicular Technology Conference (VTC2019-Spring), Kuala Lumpur, Malaysia, pp. 1-6, 2019. doi: 10.1109/VTCSpring.2019.8746407

[45] A. Bairagi, M. Munir, S. Abedin, & C. Hong. Utilization of Unlicensed Spectrum in Effective Coexistence of eMBB and uRLLC. 2018.

[46] M. Alsenwi, N. Tran, M. Bennis, S. Pandey, A. Bairagi, & C. Hong. Intelligent Resource Slicing for eMBB and URLLC Coexistence in 5G and Beyond: A Deep Reinforcement Learning Based Approach. IEEE Transactions on Wireless Communications. 2020.

[47] J. Tang, B. Shim, T. Quek. Service Multiplexing and Revenue Maximization in Sliced C-RAN Incorporated with URLLC and Multicast eMBB. IEEE Journal on Selected Areas in Communications. PP. 1-1, 2019. 10.1109/JSAC.2019.2898745.

[48] T. Ma, Y. Zhang, F. Wang, D. Wang, & D. Guo. Slicing Resource Allocation for eMBB and URLLC in 5G RAN. Wireless Communications and Mobile Computing. 1-11,2020. 10.1155/2020/6290375.

[49] J. Martyna. Reliability Enhancement of URLLC Traffic in 5G Cellular Networks. In: Gaj P., Gumiński W., Kwiecień A. (eds) Computer Networks. CN 2020. Communications in Computer and Information Science, vol 1231, 2020. Springer, Cham. https://doi.org/10.1007/978-3-030-50719-0_7

[50] S. R. Pandey, M. Alsenwi, Y. K. Tun and C. S. Hong, "A Downlink Resource Scheduling Strategy for URLLC Traffic," 2019 IEEE International Conference on Big Data and Smart Computing (BigComp), Kyoto, Japan, pp. 1-6, 2019. doi: 10.1109/BIGCOMP.2019.8679266.

[51] M. Alsenwi, S. R. Pandey, Y. K. Tun, K. T. Kim and C. S. Hong, "A Chance Constrained Based Formulation for Dynamic Multiplexing of eMBB-URLLC Traffics in 5G New Radio," 2019 International Conference on Information Networking (ICOIN), Kuala Lumpur, Malaysia, pp. 108-113, 2019. doi: 10.1109/ICOIN.2019.8718159.

[52] R. Jain, W. Hawe, D. Chiu, “A Quantitative measure of fairness and discrimination for resource allocation in Shared Computer Systems,” DEC-TR-301, September 26, 1984

[53] G. Miao, J. Zander, K. W. Sung, and S. Ben Slimane, Fundamentals of Mobile Data Networks. Cambridge: Cambridge University Press, 2016.

[54] CVX: Matlab Software for Disciplined Convex Programming. (n.d.). Retrieved from: http://cvxr.com/cvx

[55] E. Khorov, A. Krasilov and A. Malyshev, "Radio resource scheduling for low-latency communications in LTE and beyond," 2017 IEEE/ACM 25th International Symposium on Quality of Service (IWQoS), Vilanova i la Geltru, 2017, pp. 1-6, doi: 10.1109/IWQoS.2017.7969162.