Optimized Scheduling of Ultra-Reliable Low-Latency Communications Traffic for 5G Networks

Mohamed Yacine Lezzar

A Thesis
in
The Department
of
Electrical and Computer Engineering

Presented in Partial Fulfillment of the Requirements
for the Degree of Master of Applied Science (Electrical Engineering) at
Concordia University
Montreal, Quebec, Canada

November 2020

CONCORDIA UNIVERSITY
SCHOOL OF GRADUATE STUDIES

This is to certify that the thesis prepared

By: Mohamed Yacine Lezzar

Entitled: Optimized Scheduling of Ultra-Reliable Low-Latency Communications Traffic for 5G Networks

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

Master of Applied Science (Electrical and Computer Engineering)

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

Signed by the final examining committee:

______________________________ Chair
Dr. D. Qiu

______________________________ External Examiner
Dr. C. Assi (CIISE)

______________________________ Internal Examiner
Dr. D. Qiu

______________________________ Supervisor
Dr. M.K. Mehmet Ali

Approved by: ______________________________
Dr. Y.R. Shayan, Chair
Department of Electrical and Computer Engineering

______________ 20___

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

Abstract

Optimized Scheduling of Ultra-Reliable Low-Latency Communications Traffic for 5G Networks

Mohamed Yacine Lezzar

The increasingly ubiquitous applications of Ultra-Reliable Low-Latency Communications (URLLC) require innovative solutions that can only be achieved through a flexible communication system such as the The Fifth Generation (5G) New Radio (NR). Recent studies on the resource allocation for URLLC have proposed the Grant-Free (GF) scheduling instead of the traditional high latency Grant-Based (GB) scheduling, adopted in 4G Long Term Evolution (LTE). Although the GF scheduling over shared resources offers reduced latency, the possibility of achieving the reliability requirement of URLLC may be compromised due to the increased likelihood of collisions. Therefore, we propose a solution for the uplink transmissions that is capable of realizing the reliability requirement in compliance with URLLC's stringent latency budget.

The main strategy of the proposed solution is to transmit multiple uplink copies of the same packet, utilizing both dedicated and shared resources. In order to avoid additional delays, retransmissions are carried out independent of the conventional feedback from the Base Station (BS). Therefore, each packet is transmitted a pre-determined number of times, resulting in a fixed latency value for packets in the network. The network considered in this study consists of users with both periodic and sporadic traffic. Users in the network are grouped into classes according to their packet generation probabilities. Classes with high packet generation rates are characterized as periodic-traffic classes, while sporadic-traffic classes have low generation rates.

Users gain access to the available resources in the network via three different scheduling schemes. While all users access shared resources through GF scheduling, access to dedicated resources is done in two different ways, namely, Periodic Scheduling (PS) and GB scheduling. To avoid under-utilization of resources, the PS scheme is only assigned for users with high packet generation rates, while sporadic-type users access dedicated resources through the GB scheme. Although recent studies were disinclined towards the GB scheme due to its high latency, we show that the exploitation of 5G NR’s new scalable numerology results in significant reductions to GB’s latency, making it suitable for the URLLC use case. Following this latency examination, we present probabilistic expressions representing the reliability of our proposed solution.

The main contribution of this thesis to the available literature of URLLC is the presented system optimization. We optimize the system’s performance in terms of minimizing the required bandwidth or maximizing the supported traffic capacity, while satisfying the reliability requirements. Optimal performance of the system is achieved through determining the optimal allocation of resources between the considered scheduling schemes, as well as the optimal classification of user classes in the network as periodic-type or sporadic-type classes. In addition, we find the optimal packet length (for a fixed number of information bits) that results in the minimum amount of bandwidth required.

Acknowledgments

In the name of God, the Most Compassionate, the Most Merciful, all praises and thanks be to Him.

I would like to thank my supervisor, Prof. Mustafa Mehmet Ali for offering me the opportunity to conduct this work and for his continuous support, patience, and guidance throughout this program.

My deepest gratitude and love to my parents, Said and Djamila, and to my sisters, Cherine, Lilia, and Rym for their prayers, love, and encouragement during this journey. My family has a great deal of credit on everything I have accomplished so far and everything I’m yet to accomplish.

Finally, special thanks to my friends especially Abubakr Banat who always found a way to relieve the stress in any situation through humor and positive mindset.

Contents

List of Figures

List of Tables

List of Abbreviations

3GPP	The 3rd General Partnership Project
ACK	Acknowledgement
AR	Augmented Reality
AWGN	Additive White Gaussian Noise
BLER	Block Error Rate
BS	Base Station
CC	Chase Combining
DT	Diversity Transmission
Gb	Guard band
GB	Grant-Based
GF	Grant-Free
HARQ	Hybrid Automatic Repeat Request
LTE	Long Term Evolution

MCS	Modulation & Coding Scheme
MUD	Multi-User Detection
NACK	Negative Acknowledgment
NOMA	Non-Orthogonal Multiple Access
NR	New Radio
NUM1	Numerology Setting 1
NUM2	Numerology Setting 2
OFDM	Orthogonal Frequency Division Multiplexing
OMA	Orthogonal Multiple Access
OS	Orthogonal Frequency Division Multiplexing Symbol
PGF	Probability Generating Function
PMF	Probability Mass Function
PS	Periodic Scheduling
PUSCH	Physical Uplink Shared Channel

QAM	Quadrature Amplitude Modulation
QPSK	Quadrature Phase Shift Keying
RB	Resource Block
RE	Resource Element
RU	Resource Unit
SCS	Subcarrier Spacing
SG	Scheduling Grant
SIC	Successive Interference Cancellation
SNR	Signal to Noise Ratio
SR	Scheduling Request
TF	Time-Frequency
TTI	Transmission Time Interval
URLLC	Ultra-Reliable Low-Latency Communications
VR	Virtual Reality

List of Symbols

β	Signal to Noise Ratio
Δf	Subcarrier spacing
ϵ	Error probability due to channel imperfections
η	Spectral efficiency of an MCS
μ	Scaling factor for the SCS
σ_ℓ	Bernoulli trial parameter that an inactive class ℓ user generates a packet
τ_ℓ	Number of slots that a class ℓ user spends in the inactive state
ϕ_{BW}	Bandwidth requirement of one RU
ϕ_{RB}	Number of required RBs to transmit an L-sized packet
ϕ_{sym}	Number of required OFDM symbols to transmit a n-sized packet
ψ_{RB}	Number of available RBs in the channel
B_{gb}	Number of new packets generated by inactive users of GB classes
B_{gf}	Number of new packets generated by inactive users of all classes
b_ℓ	Number of packets generated by inactive class ℓ users
C	Channel capacity
F	Frame period
I_{MCS}	Modulation and Coding Scheme index
K	Number of GF repetitions over shared resources
k	Number of information bits to be transmitted (before encoding)
L	Number of user classes in the system
L_p	Size of transmitted packet in symbols
l_p	Size of transmitted packet in bits

$\boldsymbol{\ell'}$	Dividing class index between PS and GB users
$\boldsymbol{M}$	Number of available RUs in the channel
$\boldsymbol{M_{gb}}$	Number of allocated RUs for GB transmission
$\boldsymbol{M_{gf}}$	Number of allocated RUs for GF repetitions
$\boldsymbol{M_{ps}}$	Number of allocated RUs for PS transmission
$\boldsymbol{N}$	Number of users in the network
$\boldsymbol{N_{gb}}$	Number of users allocated to the GB scheme
$\boldsymbol{N_{ps}}$	Number of users allocated to the PS scheme
$\boldsymbol{n_{\ell}}$	Number of class ℓ users in the system
$\boldsymbol{P_b}$	Probability of blocking in $K+1$ slots
$\boldsymbol{P_{co}}$	Probability of collision in the GF scheme
$\boldsymbol{P_S}$	Probability that a GF transmission of a packet will be successful
$\boldsymbol{Q}$	Gaussian cumulative distribution function
$\boldsymbol{Q_m}$	Modulation order
$\boldsymbol{q_{gb}}$	Probability of failure of one GB transmission plus K GF repetitions
$\boldsymbol{q_{ps}}$	Probability of failure of one PS transmission plus K GF repetitions
$\boldsymbol{p_{gb,\ell}}$	Probability that a GB-based class ℓ user will have a GB packet to transmit during a slot
$\boldsymbol{p_{gf,\ell}}$	Probability that a class ℓ user will have a GF packet to transmit during a slot
$\boldsymbol{R}$	Coding rate
$\boldsymbol{T}$	Traffic capacity of the system
$\boldsymbol{t_{TTI}}$	Duration of Transmission Time Interval
$\boldsymbol{V}$	Channel dispersion
$\boldsymbol{W}$	Channel bandwidth in Hz
$\boldsymbol{W_{gb}}$	Amount of bandwidth in Hz allocated for GB

$\boldsymbol{W_{gf}}$	Amount of bandwidth in Hz allocated for GF
$\boldsymbol{W_{ps}}$	Amount of bandwidth in Hz allocated for PS
$\boldsymbol{w_{RB1}}$	Amount of bandwidth needed in Hz for a single RB of NUM1
$\boldsymbol{w_{RB2}}$	Amount of bandwidth needed in Hz for a single RB of NUM2

Chapter 1

Introduction and Literature Review

1.1 Chapter Overview

This chapter commences the thesis by presenting the studied problem of URLLC, proposing a solution for the problem, and comparing the proposed solution with the existing literature. Such comparison allows for the recognition of this work's contribution to the use case of URLLC. At the end of the chapter, a brief mention of the contents of the thesis is given.

1.2 Ultra-Reliable Low-Latency Communications

1.2.1 Services of 5G NR

The Fifth Generation (5G) New Radio (NR) introduces three new use cases (services): enhanced Mobile Broadband (eMBB), massive Machine-Type Communications (mMTC), and Ultra-Reliable Low-Latency Communications (URLLC) [1] [2]. Features of the three new services are summarized in Fig. 1.1. The most challenging use case is URLLC due to its extremely stringent requirements presented in the following subsection.

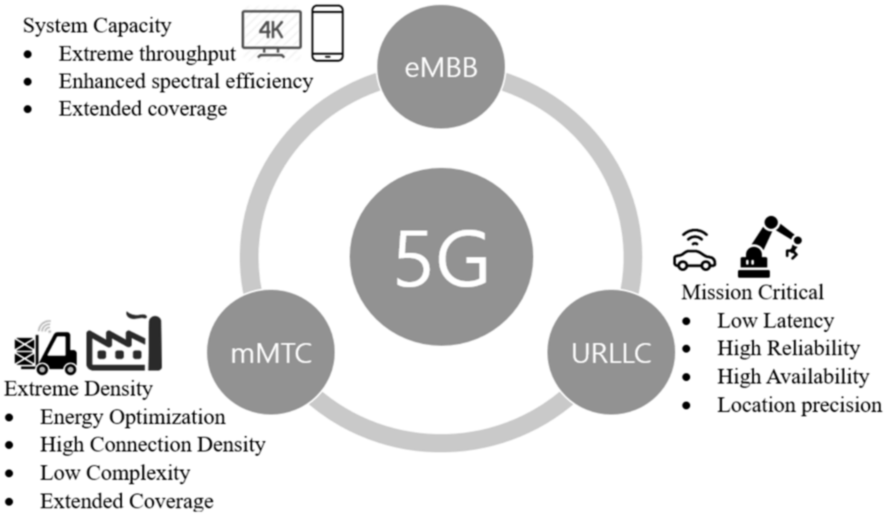

Fig. 1.1. Services of 5G NR [3].

1.2.2 Requirements of URLLC

According to The 3rd Generation Partnership Project (3GPP), a general URLLC requirement for a short packet (usually 32 bytes) is 99.999% success rate reliability within 1 ms latency including possible retransmissions [1]. Traditionally, a packet success rate of 99.999% can be achieved through conventional retransmission techniques such as Hybrid Automatic Repeat Request (HARQ). However, the conflicting requirement of tight latency bounds in the scale of milliseconds makes it a very challenging task to accomplish. These URLLC requirements are essential to various emerging applications mentioned briefly in the following subsection.

1.2.3 Applications of URLLC

The service of URLLC enables real-time control processes, making it essential for a lot of mission-critical applications such as: Virtual/Augmented Reality (VR/AR) [4], factory automation [5] [6], and intelligent transportation, which facilitates the implementation of automated driving by mitigating machine errors while limiting response delays [7]. Another revolutionary application that depends on URLLC is the medical service of tele-surgery [8]. Tele-surgery enables the process of remote surgical consultations as well as remote surgery, which would make substantial advances in the medical field. Other potential applications of URLLC include smart grids, fault detection, and Tactile Internet (TI).

1.3 Proposed Solution

This thesis proposes a resource scheduling and allocation solution for a heterogenous network, supporting URLLC-type users with both periodic and sporadic traffic characteristics. Although a similar solution could be applied for downlink, our proposal is focused on the uplink traffic. In the network, each active user (has a packet to transmit) sends an uplink transmission once through dedicated resources and K times over a shared pool of resources. Therefore, to ensure meeting the reliability requirement of URLLC, a user transmits the same packet a total of $K + 1$ times. We consider three scheduling schemes that enable users to access the available resources, namely, Periodic Scheduling (PS), Grant-Based (GB) scheduling and Grant-Free (GF) scheduling. The first two allow users to access the dedicated resources, while GF is the scheme assigned to accessing the shared pool.

The proposed solution suggests that users with sporadic traffic are allocated dedicated resources on request through GB. On the other hand, pre-allocated dedicated resources are provided to users offering high loads (periodic traffic) through PS. The periodic scheduling of resources is prone to wastage of resources in case a user does not have a packet to transmit during the allocated slot. However, it is still a better choice than the GB scheme for users with high packet generation probability. Since dedicated resources are collision-free, the probability of a successful transmission using GB or PS is very high. The K supplementary repetitions are chosen to be on GF shared resources for two reasons: 1) to meet with the strict latency budget; 2) to avoid wasting dedicated resources in case the transmission through GB or PS was successful.

1.4 Related Work

In this subsection we describe the related work in the literature. In [9], three GF transmission schemes are considered, namely, Reactive, K-Rep and Proactive scheme. In the Reactive scheme, the user sends an uplink transmission and waits for the Base Station (BS) to send a feedback in the form of positive or negative acknowledgement. Retransmission of the same packet occurs when the users receives a negative acknowledgment from the BS. In the K-Rep scheme, the user transmits the same packet K times to the BS without waiting for a feedback, resulting in a reduced delay compared to the Reactive scheme. Lastly, the Proactive scheme is similar to the K-Rep scheme but with the possibility of terminating the transmission process upon receiving a positive acknowledgement even if the K configured repetitions are not completed. The performance of the three GF schemes is evaluated through Monte Carlo system level simulation and compared against the baseline GB transmission as a function of the load. Note that the GB scheme considered in [9] for comparison is similar to the Reactive scheme with the difference of the grant needed before the transmission. Results show that GF has overall better latency performance than GB, especially the Proactive scheme. It is also shown that among the GF schemes, the Reactive one can support the largest load (400 packets per second per cell) and shares the lead with GB as the most resource efficient.

In [10], a HYBRID transmission scheme is proposed based on a combination of the NACK-based and BLIND schemes, which are other names of the Reactive and Proactive schemes [9], respectively. In the HYBRID scheme, transmission starts as NACK-based but switches to BLIND in two cases: 1) after N NACK-based transmissions; 2) at the last transmission slot before reaching the latency deadline. Moreover, Chase Combining (CC)-HARQ is implemented which allows the receiver to combine all received transmission attempts. The performance of the three schemes (NACK, BLIND, and HYBRID) is compared through simulation. In the simulation, retransmission delays are considered, while the delay due to the scheduling request of the first transmission is assumed to be zero. It is also assumed that feedback from BS is always received error-free. Results show that the NACK-based model has the best spectral efficiency (i.e. the amount of information transmitted over a given bandwidth). However, it was unable to meet the latency budget for the required reliability target. The BLIND scheme shows the best results in terms of latency and reliability but it has poor spectral efficiency due to the unnecessary transmissions before reception of ACK. Finally, the HYBRID model was able to satisfy both latency and reliability requirements with spectral efficiency that is close to the NACK scheme.

Another hybrid scheme is proposed in [11], based on multiple grant-free transmissions on the shared pool and/or dedicated resources allocated periodically. A probabilistic model is derived to determine the amount of resources required and the number of repetitions needed. The analytical model considers the following two cases: 1) Repeated transmissions without early termination (K -rep); 2) Repeated transmissions with possible early termination (Proactive). The proposed scheme is compared with a conservative scheme in which a user is always allocated a dedicated resource. In terms of resource utilization, results prove the superiority of the proposed scheme over the conservative one. It can also be seen that the scheme using repetitions with early stop results in slightly better resource utilization than the one without early stop. Moreover, in poor channel conditions, the gain in resource utilization is significant when the traffic load is low. However, the gain decreases as the traffic load increases, due to the allocation of most of the resources as dedicated resources.

In [12], two collision reduction techniques are considered for grant-free transmissions, namely, Diversity Transmission (DT) and Multiuser Detection (MUD). DT is performed by multiple transmissions of the same packet in a slotted ALOHA fashion where the resource blocks used in each consecutive transmission are randomly chosen. While, MUD is done by assuming the receiver has advanced capabilities where it can successfully decode a collided packet with received power that is at least 5 dB stronger than the interfering packets. The probability of a successful

MUD is obtained via simulation, while collision probability is provided analytically without considering the channel impact on the final reliability. Moreover, the discussed model here presumes sporadic transmissions only. Comparisons are made with default multi-channel slotted ALOHA access scheme and results from both simulation and analysis discuss the supported population for different collision probabilities and arrival rates. It is shown that the analytical model breaks down at high arrival rates. Overall, the two proposed techniques show improvement in reliability, especially if combined together, however, the multi-user detection require relatively complex implementation and may result in increased latency (which has not been studied). Other works proposing similar variations of grant-free scheduling techniques for URLLC have been presented in [13] - [15].

The authors of [16] introduced the non-orthogonal multiple access hybrid automatic repeat request (NOMA-HARQ) scheme, which allows a user requiring retransmission to be scheduled on the same TF-block (Time-Frequency block) with another user transmitting a new packet, in order to save resources. Preceding an uplink transmission, the user receives a scheduling message from the BS with information about the allocated TF-block and the optimal transmission power. The assigned power to each user is the minimum power needed to guarantee the reliability requirement. It is assumed that each packet can be retransmitted at most two times, however, latency analysis is not considered. It is also assumed that the receiver has optimal Successive Interference Cancellation (SIC) capabilities. Error probability is derived analytically along with system-level simulations to compare with baseline orthogonal multiple access (OMA) transmission scheme on dedicated resources. Simulation results show that NOMA-HARQ can offer gains in terms of spectral efficiency and proves to work best in low error rate conditions. Similar proposals of NOMA for URLLC were discussed in [17] and [18].

The Proactive scheme presented in [9] can trigger early stop of repetitions only after at least three slots (TTIs) from receiving a successful packet. This is due to processing time at the BS, the time taken to transmit an ACK and process it. Therefore, at least three unnecessary repetitions will be transmitted. Incidentally, in our proposed solution, the maximum number of GF repetitions that can be carried out by GB-based users before reaching the $0.6\,ms$ latency deadline is three repetitions, as will be shown in our latency analysis. Thus, we choose in our analysis $K = 3$ which makes our proposed GF repetitions similar to the Proactive scheme of [9] in terms of resource efficiency. As for the Reactive scheme also presented in [9] and [10], it is the most resource efficient since retransmission occur only when needed. However, waiting for feedback from the receiver before retransmitting is not an optimal choice for URLLC applications as it can easily result in exceeding the latency budget. The hybrid scheme proposed in [11] is very similar to our proposed PS scheme with GF repetitions, which is only suitable for users with periodic-type traffic. The GF K-repetition model in [12] might not be sufficient to meet the reliability target in the cases of a high load system or users with periodic traffic, even with the support of advanced receiver techniques. Finally, the NOMA-HARQ scheme in [16] is the best in terms of resource efficiency since two users can utilize the same resource simultaneously. Although their work didn't include latency analysis, the significant delay resulting from scheduling message, feedback, and the SIC processing time can be anticipated.

1.5 Contribution of the Thesis

A resource-efficient scheduling solution that supports URLLC use cases is studied in this thesis. The principal contributions of this thesis are:

- To provide a resource scheduling solution that can support both periodic and sporadic traffic, while satisfying the requirements of URLLC.
- To demonstrate a way that makes the Grant-Based scheduling feasible for URLLC applications in terms of latency, through the use of scalable numerology.
- To present specific details on the delay components in each of the considered scheduling schemes. These delay components influence the choices in designing the proposed transmission protocols, in order to meet the latency budget.
- To derive the failure probability of a transmission in each of the scheduling schemes, through probabilistic analysis. These probabilities are considered as a measure of the system's reliability.

- To classify the existing users in the system into different classes based on their packet generation rate. Then find the optimal position of the dividing line that separates between periodic and sporadic classes.

- To find the optimal allocation of resources between the three considered scheduling schemes that would result in either: 1) minimizing the amount of bandwidth required to support a certain number of users; or 2) maximizing the supported traffic capacity for a given amount of bandwidth. This is achieved by determining the optimal dividing class index which acts as a separation line between PS-based and GB-based classes, then finding the minimum number of resource units required in each scheduling scheme to achieve the target reliability.

To the best of our knowledge, this is the first work on optimization of the resource allocation for multi-class URLLC systems. We believe that these contributions would advance the field of mission-critical communications as they provide considerable additions to the work available in the literature.

1.6 Structure of the Thesis

This thesis contains five chapters and the remainder is organized as follows:

- Chapter 2 – In this chapter, we explore the background information related to the physical layer of 5G NR, which are essential for the reliability and latency analyses of the proposed solution. We explain the structure of the physical resources in order to obtain the requirement for transmitting a fixed length packet. Further, we introduce a new feature of 5G called ‘scalable numerology’, and provide two different numerology settings to be used throughout this thesis. Moreover, we show the relations leading to the error probability of a finite-length packet used in literature.

- Chapter 3 – This chapter is initiated by describing the network model assumed in this thesis. It also introduces the three scheduling schemes by explaining the mechanism and objective of each one of them. Next, we present the reliability and latency analyses of the three scheduling schemes. Latency analysis allows us to prove the feasibility of the proposed solution in compliance with the latency requirements of URLLC. While the reliability analysis will be pivotal in Chapter 4, as the presented probabilities will be the main constraints of the system optimization.

- Chapter 4 – In this chapter, we optimize various parameters of the system according to different assumptions and scenarios. This optimization allows for the enhancement of the system's performance while making sure that URLLC's requirements are being met. The overall performance of the system is evaluated through either the bandwidth required or the traffic capacity. The objective is to minimize the amount of bandwidth needed to support a given number of users, or maximize the traffic capacity for a given amount of bandwidth. All cases and scenarios are subject to the constraints of the URLLC's stringent requirements.
- Chapter 5 – A conclusion of the thesis along with some possible future work are included in this final chapter.

Chapter 2

Physical Layer in 5G

2.1 Introduction

In this chapter, we provide the necessary background information related to the physical layer of 5G NR. Although physical layer aspects are not the main scope of this thesis, some information are needed for the reliability and latency analyses of the proposed solution. This chapter explains the composition of the physical resources in terms of both frequency and time components. Moreover, it introduces the new scalable numerology feature of 5G, which is the main enabler to the proposed grant-based scheduling scheme. Finally, we show the effects of channel coding and channel conditions on the probability of successfully decoding a packet.

2.2 Physical Resources

In 5G New Radio (NR), the signal's waveform is based on the Orthogonal Frequency Division Multiplexing (OFDM). The smallest physical frequency resource which consists of one OFDM subcarrier is called a Resource Element (RE) and each subcarrier carries one symbol of data. A Resource Block (RB) consists of 12 consecutive OFDM subcarriers that are spaced apart in frequency by the Subcarrier Spacing (SCS), denoted Δf. Letting ϕ_{RB} denote the number of RBs used for each transmission, then the number of symbols that can be transmitted simultaneously is given by $\phi_{RB} \times 12$ during each OFDM symbol transmission time. Packet transmissions have been divided into Transmission Time Intervals (TTIs), where ϕ_{sym} denotes the number of OFDM symbol transmission times in a TTI. The number of transmitted symbols per TTI, L_p, is given by,

$$L_p = \phi_{sym} \times \phi_{RB} \times 12 \ \text{(symbols/TTI)} \tag{2.1}$$

while the number of transmitted bits per TTI, l_p, is given by,

$$l_p = L_p \times Q_m \quad \text{(bits/TTI)} \tag{2.2}$$

where Q_m is the modulation order (in bits/symbol). Possible values of Q_m in 5G NR are listed in Table 2.1.

Table 2.1. Order of various modulation schemes in bits/symbol.

Modulation Scheme	**Modulation Order (Q_m)**
QPSK	2
16 QAM	4
64 QAM	6

Let ϕ_{BW} denote the bandwidth requirement of ϕ_{RB} resource blocks,

$$\phi_{BW} = 12 \times \Delta f \times \phi_{RB} \tag{2.3}$$

Table 2.2. Summary of notations in Chapter 2.

Notation	**Definition**
Δf	Subcarrier Spacing
RB	A resource block, it consists of 12 subcarriers
ϕ_{RB}	Number of RBs per transmission
$\phi_{BW} = 12\Delta f \phi_{RB}$	Bandwidth requirement of ϕ_{RB} blocks
TTI	Transmission time interval
ϕ_{sym}	Number of symbols per subcarrier per TTI interval
$L_p = 12\phi_{sym}\phi_{RB}$	Total number of transmitted symbols per TTI by ϕ_{RB} resource blocks
$l_p = L_p Q_m$	Number of bits transmitted per TTI

Let us define a Resource Unit (RU) as a frequency chunk consisting of ϕ_{RB} RBs needed to transmit a packet of length L_p symbols. A channel that has ψ_{RB} available Resource Blocks, the number of available Resource Units M in that channel is given by,

$$M = \psi_{RB}/\phi_{RB} \tag{2.4}$$

where each packet requires ϕ_{RB} resource blocks for its transmission. In the following, it will be assumed that packet lengths are L_p symbols, thus each packet will require one RU for its transmission and packet transmission will take one TTI.

2.3 Numerology

One of the most significant changes introduced in 5G NR is the scalable numerology. For the case of low latency applications, numerology enables transmissions of shorter duration by either increasing the SCS (which reduces the duration of an OFDM symbol) or by using mini-slots consisting of 2, 4, or 7 OFDM symbols, instead of the standard 14 OFDM transmission slot. The values of the SCS, Δf, are scaled from the fundamental 15 KHz SCS as,

$$\Delta f = 2^{\mu} \times 15 \text{ KHz} \tag{2.5}$$

where $\mu \in \mathbb{Z} : \mu \in [1,5]$ is the scaling factor [19]. In our study, we will be examining two different numerology settings to transmit an L_p-sized packets, namely, NUM1 and NUM2, as shown in Table 2.3. The parameters in the table assume the maximum packet length to be $L_p = 168$ symbols. Note that t_{TTI} in the table refers to the duration of each Transmission Time Interval (TTI). Visual representation of the two numerology settings is shown in Fig. 2.1, where each row corresponds to an RB and each column corresponds to an OFDM symbol. Note that in the remainder of this thesis we assume the values of Δf and ϕ_{sym} listed in Table 2.3 for NUM1 and NUM2, while the value of ϕ_{RB} will depend on the packet length L_p, which can be calculated using (2.1).

Table 2.3. Numerology settings.

Numerology	$\boldsymbol{\Delta f}$	$\boldsymbol{\phi_{sym}}$	$\boldsymbol{t_{TTI}}$	$\boldsymbol{\phi_{RB}}$	$\boldsymbol{\phi_{BW}}$
NUM1	15 KHz	2 symbols	142.8 μs	7 RBs	1260 KHz
NUM2	60 KHz	7 symbols	125 μs	2 RBs	1440 KHz

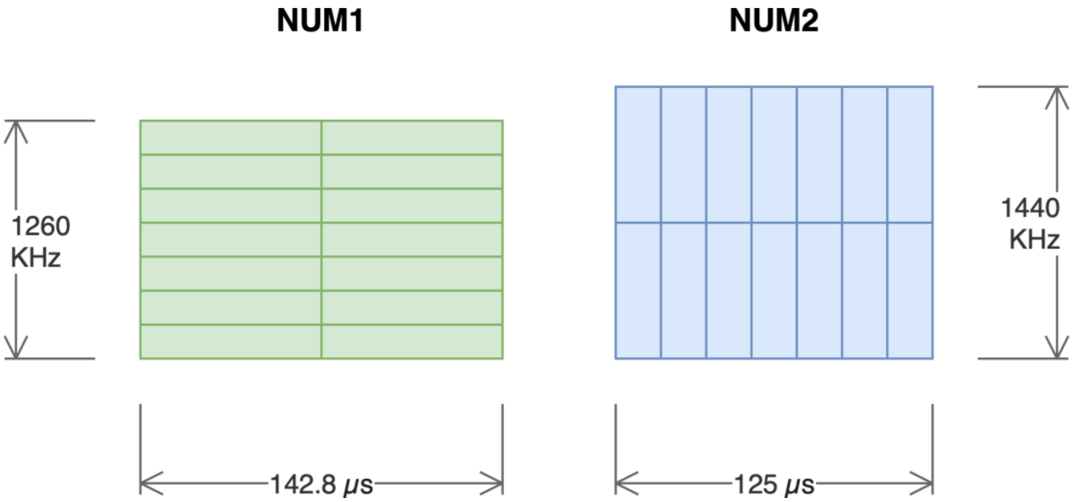

Fig. 2.1. NUM1 and NUM2 in time-frequency plane.

Table 2.4 presents the number of available RBs, ψ_{RB}, for different channel bandwidths and SCS values according to 3GPP's specifications [20]. Each channel will have a guard band (Gb) with the adjacent channels to avoid interference. Each column in the table corresponds to a channel bandwidth and each batch of three rows correspond to a SCS value. The first row in each batch gives ψ_{RB} for that channel, second row gives the Gb between channels, and the third row is the yield (Y) of the channel. The yield is the useful bandwidth of the channel after subtraction of the Gb on each edge of the channel.

We can determine the number of resource units, M, that each channel supports for both numerologies from (2.4), where ϕ_{RB} and ψ_{RB} are taken from Table 2.3 and Table 2.4, respectively. If we take a 20 MHz channel bandwidth as an example, we will have $M = 15$ for the case of NUM1, while for NUM2 the number of available RUs will be $M = 12$. In this case it is obvious that NUM1 can support more traffic than NUM2 because it can have more RUs. The fact that wider SCS uses more bandwidth, hence, fewer RBs per channel is the basis of our comparison between the two numerology settings. However, a wider SCS numerology has the upper hand in terms of latency, which will be discussed in section 3.4

Table 2.4. Number of RBs per Channel Bandwidth.

			Channel Bandwidth											
μ	Δf		5 MHz	10 MHz	15 MHz	20 MHz	25 MHz	30 MHz	40 MHz	50 MHz	60 MHz	70 MHz	80 MHz	90 MHz
0	**15 KHz**	ψ_{RB}	25	52	79	106	133	160	216	270	N/A	N/A	N/A	N/A
		Gb [KHz]	242.5	312.5	382.5	452.5	522.5	592.5	552.5	692.5				
		Y [MHz]	4.5	9.4	14.2	19.1	23.9	28.9	38.9	48.6				
1	**30 KHz**	ψ_{RB}	11	24	38	51	65	78	106	133	162	189	217	245
		GB [KHz]	505	665	645	805	785	945	905	1045	825	965	925	885
		Y [MHz]	4	8.6	13.7	18.4	23.4	28.1	38.2	47.9	58.3	68	78.1	88.2
2	**60 KHz**	ψ_{RB}	N/A	11	18	24	31	38	51	65	79	93	107	121
		GB [KHz]		1010	990	1330	1310	1290	1610	1570	1530	1490	1450	1410
		Y [MHz]		7.9	13	17.3	22.3	27.4	36.7	46.8	56.9	67	77	87.1

2.4 Error Probability in Finite Blocklength

One of the main challenges of wireless communications is that signals are prone to distortions due to channel imperfections such as noise and fading, as well as interference from other active devices in the network. To achieve reliable communications over unreliable channels, error detection and correction techniques are used by adding extra (redundant) data to the transmitted information. The process of adding redundant data and then recovering the information bits at receiving end, shown in Fig. 2.2, is called ‘channel coding’. At the transmitter side, an encoder maps k information bits into L_p symbols referred to as the ‘blocklength’ (or packet length).

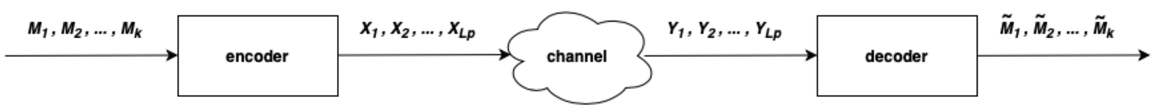
$M_1, M_2, \ldots, M_k$
encoder
$X_1, X_2, \ldots, X_{Lp}$
channel
$Y_1, Y_2, \ldots, Y_{Lp}$
decoder
$\tilde{M}_1, \tilde{M}_2, \ldots, \tilde{M}_k$

Fig. 2.2. Channel coding.

In coding theory, the ratio $R = k/L_p$ in bits per symbol is called the 'coding rate'. The objective is to design codes that would maximize R while minimizing the packet error rate ϵ, also referred to as Block Error Rate (BLER). A fundamental result is Shannon's Channel Capacity C which gives an upper bound to the achievable rate while maintaining a very small ϵ [21]. However, Shannon's capacity assumes very large values of L_p (long packets), therefore, it is not suitable for applications with finite blocklength (short packets) such as URLLC. For such cases, it is shown in [22] that the maximal achievable coding rate R with error probability ϵ in the finite blocklength regime can be expressed as,

$$R = C - \sqrt{\frac{V}{L_p}} Q^{-1}(\epsilon) + \mathcal{O}\left(\frac{\log L_p}{L_p}\right) \tag{2.6}$$

where V is the channel dispersion, $\mathcal{O}\left(\frac{\log L_p}{L_p}\right)$ denotes the remainder of terms of order ${}^{\log L_p}\!/_{L_p}$ and Q is the Gaussian cumulative distribution function given by,

$$Q(x) = \int_{x}^{\infty} \frac{1}{2\pi} e^{-t^2/2}\, dt \tag{2.7}$$

In our analysis, we will be using the normal approximation of (2.6) which was found to be a good estimation of the term $\mathcal{O}\left(\frac{\log L_p}{L_p}\right)$, especially for blocklengths higher than 200 symbols [22],

$$R \approx C - \sqrt{\frac{V}{L_p}} Q^{-1}(\epsilon) + \frac{1}{2L_p} \log L_p \tag{2.8}$$

Solving for ϵ yields the maximal probability of packet error rate for a given coding rate $R = {}^{k}/_{L_p}$ as,

$$\epsilon \approx Q\left(\frac{L_p C - k + (\log L_p)/2}{\sqrt{L_p V}}\right) \tag{2.9}$$

For a real additive white Gaussian Noise (AWGN) channel with a Signal to Noise Ratio $SNR = \beta$, capacity C and channel dispersion V can be calculated as [22],

$$C = \frac{1}{2}\log(1+\beta) \tag{2.10}$$

$$V = \frac{\beta}{2}\frac{\beta+2}{(\beta+1)^2}\log^2 e \tag{2.11}$$

Note that channel capacity here is measured in bits per input symbol into the channel (bits per channel use), for instance, if the capacity $C = 1$, it means that each transmitted symbol contains one bit.

2.5 Conclusion

In this chapter, we introduced the time-frequency resources in 5G NR, which are based on the OFDM waveform. We gave the expressions relating these resources with the number of bits transmitted in a packet. These expressions allow us to evaluate the bandwidth requirement for transmitting an L_p-sized packet. Further, we explained the feature of scalable numerology and showed its effect on the resources, especially in terms of bandwidth. Moreover, we presented two numerology settings, namely, NUM1 and NUM2, to be used in the remainder of this thesis. Finally, we explained the relations leading to obtaining an expression for the packet error rate in the context of finite blocklength, derived in [22].

Chapter 3

System Model and Analysis

3.1 Introduction

This chapter lays the fundamental assumptions of the network model considered in this thesis. It also explains the mechanism of the three foundational scheduling schemes in which the proposed solution is built around. In addition, we examine the latency of each of these scheduling schemes. Following that, the probabilistic analysis is presented according to the assumptions of the model. Latency and reliability analyses are essential for showing the feasibility of the proposal as a valid URLLC solution.

3.2 Network Model

A single cell network with one Base Station (BS) is assumed and only uplink transmissions are considered. There are N users in the network that alternate between two states of active and inactive. In the active state a user is serving a packet and in the inactive state it's in the process of generating a new packet. Thus, at any time a user may have single packet waiting for transmission. The time axis is slotted into the TTI durations that packet transmissions take place. The available resources are divided into two categories as dedicated and shared resources. A resource is referred to as 'dedicated' when it is assigned to only one user during a time-slot (or TTI), therefore, the probability of packet collisions is zero. On the other hand, a 'shared' resource is prone to collisions since multiple users may transmit on the same resource. A user transmits a packet once through dedicated resources and K times over the shared resources. We note that each packet is allowed only a single transmission during a slot. Since a user transmits the same packet a total of $K + 1$ times, service time of a packet will take $K + 1$ slots, which will be referred to as a frame, denoted by F. The scheduling schemes that enable users to access both dedicated and shared resources are discussed in section 3.3

The objective is to achieve a packet loss probability smaller than 10^{-5} within $0.6\,ms$ latency. The transmitted packets may experience failure due to channel imperfections. The K copies of packets transmitted over shared resources, which will be referred to as 'repetitions', are used in order to help meet the stringent reliability requirement in case the transmission over dedicated-resources fails. Each transmission of a packet is assumed to occupy one RU for the duration of one time slot (or TTI). When a user is not serving a packet (idle), it may generate a single packet during a slot, however, it may not generate a new packet during the service time of an existing packet. Note that since NUM1 and NUM2 have different TTI durations as shown in Table 2.3, a time slot is assumed to have the duration of the longer TTI between the two, that is the TTI for NUM1 ($t_{TTI} = 142.8\,\mu s$).

3.3 Scheduling Schemes

In the above, it was stated that a user will have a transmission over dedicated resources and K repetitions over shared resources. The dedicated transmission may happen either through a periodically assigned slot to a user or through grant-based scheduling. A user will be assigned a dedicated slot periodically if the probability of packet generation during a slot is high, otherwise, the grant-based scheduling will be preferred for transmission over dedicated resources. Repetitions will always be transmitted over shared resources. The use of shared resources will be referred to as grant-free scheduling. The network resources will be divided among the periodic, grant-based and grant-free scheduling.

3.3.1 Periodic Scheduling

In the Periodic Scheduling (PS) scheme, each user is assigned a slot every frame to access a single RU from the M_{ps} allocated RUs for periodic dedicated transmissions. Thus, this scheme allows users to have an opportunity for one collision-free transmission for each of their packets. When users generate a packet and become active, they start by transmitting the repetitions over the shared resources until the time comes for their periodically assigned dedicated resources, assuming users are synchronized in time with the BS. In case a user does not generate a packet to transmit during a frame, the pre-allocated dedicated RU for that user would be wasted during that frame.

An example of the PS scheme is shown in Fig. 3.1, where the process is shown from the perspective of the user of interest 'U1'. Assuming the channel has $M_{ps} = 2$ periodically-scheduled RUs, referred to as D1 and D2, in addition to 4 shared RUs (S1, S2, S3 and S4) for $K = 3$ repetitions over shared resources for each packet. The user U1 is assigned the dedicated resource D1 at the first slot of each frame F. The idea of having a single collision free transmission with dedicated resources allocated periodically makes the PS scheme desirable, however, with a high number of users it might not be feasible. Evidently, the PS scheme is only valid if,

$$M_{ps} \geq \frac{N_{ps}}{K+1} \tag{3.1}$$

where M_{ps} refers to the number of periodically-scheduled RUs available per slot for dedicated transmission and N_{ps} is the number of users accessing them.

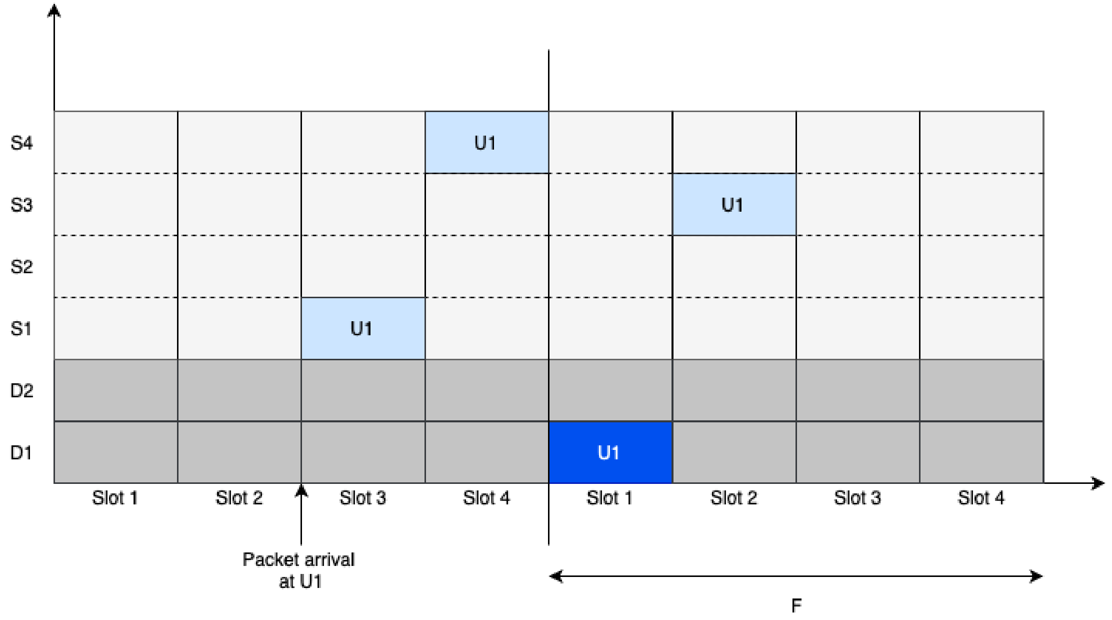

Fig. 3.1. Transmission of U1's packet through the PS scheme.

3.3.2 Grant-Based Scheduling

The Grant-Based (GB) scheduling is another scheduling scheme that allows single transmission of a packet over dedicated resources. In the GB scheme, the user has to make a request for access to dedicated resources. Fig. 3.2 shows the procedure of GB, which starts when an active user sends a Scheduling Request (SR) to the BS requesting access to dedicated resources. The BS in return sends a Scheduling Grant (SG) to the user for the next slot, granting it access to a specific RU. However, since the number of users might be greater than the number of available RUs, some users' requests could not be met in the next time slot, which results in blocking during that slot.

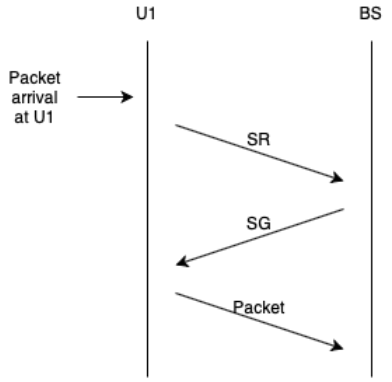

Fig. 3.2. GB procedure.

3.3.3 Grant-Free Scheduling

Unlike the two previously discussed scheduling schemes, Grant-Free (GF) scheduling provides transmissions only through shared resources. In our model, these collision-prone resources are used to transmit the K repetitions of packets by users. As the name suggests, this scheme allows users to transmit their repetitions without the need for a grant from the BS. The lack of coordination between users and the BS allows for more transmissions within a shorter time interval than the GB scheme, as it will be demonstrated in section 3.4 However, this may result in packet collisions which is why this scheme is perceived only as a supplementary method of transmission as far as reliability is concerned. Repetitions are sent in consecutive time slots unless there is an opportunity for a dedicated-resource transmission, in which case repetitions are interrupted so that the higher-priority dedicated-resource transmission can take place. Following this interruption, the GF repetitions resume only if the latency budget has not yet been exceeded. Note that in the GF scheduling, a different RU is used for each repetition (frequency hopping), this helps mitigate the impact of channel fading and collisions.

3.4 Latency Analysis

In this section, we discuss the latency overhead of each of the scheduling schemes explained earlier, and show how these schemes can achieve the latency requirement of URLLC.

3.4.1 Periodic Scheduling

Since dedicated resources are pre-allocated periodically in PS, there is no overhead latency in the process of assigning them. Therefore, the only latency component in the transmissions through PS will be the duration of TTI (Transmission Time Interval), denoted as t_{TTI}. The value of t_{TTI} is determined according to the numerology setting used. Recall from our comparison between the two proposed numerology settings in section 2.3 that NUM1 performs better in terms of the number of available RUs. Hence, NUM1 will be considered for the PS scheme, resulting in a total latency of $t_{TTI} = 142.8\ \mu s$ for one dedicated-resource transmission.

3.4.2 Grant-Based Scheduling

Recall from the previous section that the process of GB involves the exchange of a Scheduling Request (SR) and a Scheduling Grant (SG) between a user and the BS through the Physical Uplink Control Channel (PUCCH). This process results in additional overhead delay components, although it allows users to access a dedicated RU in the Physical Uplink Shared Channel (PUSCH) efficiently. Delay components of the GB scheduling procedure are listed in Table 3.1, with a brief explanation of each of them. Note that the duration components N1 and N2 (steps 3 and 5) depend on the subcarrier spacing (SCS) used, as listed in Table 3.3, with the duration represented in terms of the number of OFDM symbols (OS).

Consequently, this would influence the choice of the numerology setting to be used. Since the duration of an OFDM symbol increases in narrower SCS (as shown in Table 3.2),15 KHz and 30 KHz SCS are not viable options for GB transmission as they result in transmissions that would exceed our $0.6\ ms$ latency target. Therefore, the suitable numerology setting for the GB scheme is NUM2 due to its wider SCS (60 KHz) allowing for a faster scheduling process. In Fig. 3.3, it is shown that one GB transmission takes 31 OS which corresponds to $0.553\ ms$ when operating by NUM2. In Table 3.4, we present the latency of a GB transmission for 15 KHz, 30 KHz, and 60 KHz subcarrier spacings.

Table 3.1. GB transmission's components.

Step	Description	Explanation	Duration
1	SR preparation	Time to prepare for the transmission of the SR (including alignment time)	2 OS
2	SR	Transmission duration of the SR	1 OS
3	SR processing	Time it takes for the BS to process the SR and choose a dedicated resource	N1
4	SG	Transmission duration of the SG	1 OS
5	PUSCH preparation	The time between receiving the SG and the earliest PUSCH transmission possibility	N2
6	TTI	Number of symbol transmission times for 15/30/60KHz subcarrier spacing in a TTI interval	2/4/7 OS

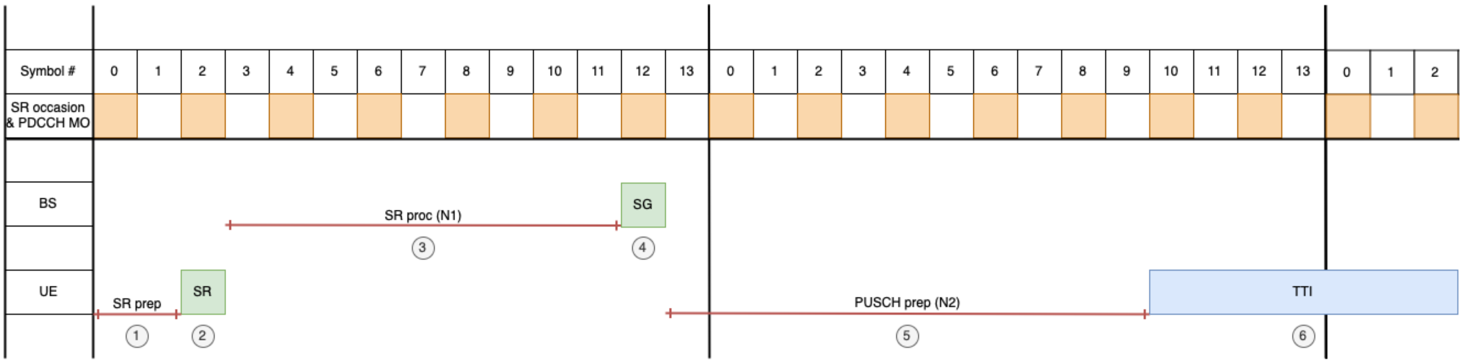

Fig. 3.3. GB transmission (60 KHz SCS).

Table 3.2. Symbol duration for various SCS.

SCS	Symbol duration (in μs)
15 KHz	71.42
30 KHz	35.71
60 KHz	17.85

Table 3.3. Variable-duration components.

SCS	N1 (in OS)	N2 (in OS)
15 KHz	3	5
30 KHz	4.5	5.5
60 KHz	9	11

Table 3.4. Latency of a GB transmission.

SCS	Latency (in OS)	Latency (in ms)
15 KHz	14	0.9988
30 KHz	18	0.64278
60 KHz	31	0.55335

Let m denotes the overhead delay imposed by the grant process mentioned in steps 1-5 in Table 3.1. Since GB operates on NUM2, the overhead delay is 24 OFDM symbols which is equivalent to $428.4\ \mu s$. As mentioned earlier, a slot has the duration equivalent to the TTI of NUM1 ($t_{TTI} = 142.8\ \mu s$). Thus, the overhead duration is equivalent to $m = \frac{428.4\ \mu s}{142.8\ \mu s} = 3$ slots.

3.4.3 Grant-Free Scheduling

Similar to the PS scheme, in GF repetitions there is no latency overhead as users transmit their uplink packets without coordination with the BS. Accordingly, NUM1 is considered for the GF repetitions, since it would result in a higher number of RUs than NUM2, ultimately reducing the probability of collisions. This yields a latency of $t_{TTI} = 142.8\ \mu s$ (one slot) per repetition out of the total K repetitions. Therefore, in case of $K = 3$ repetitions, a GB-based user can transmit the GF repetitions of its packet during the $m = 3$ slots overhead mentioned in the previous subsection.

3.5 Reliability Analysis

This section presents reliability analysis of the scheduling schemes introduced in section 3.3 . In the proposed model, each packet is transmitted $K + 1$ times, once over dedicated resources and K times over shared resources. A packet is lost if none of the $K + 1$ transmissions of a packet is successfully received by the BS. The objective of the reliability analysis is to determine the packet loss probabilities. It is assumed that each user in the inactive state generates a new packet according to an independent Bernoulli trial during a slot and then transits to the active state to serve that packet. It will be assumed that the users are divided into a number of classes according to the parameter of the Bernoulli trial. Each class of users will be classified as either a PS or GB class. Let us introduce the following notation,

L Number of classes of users.

ℓ' The dividing class index between PS and GB users. Classes $\ell \leq \ell'$ are PS classes and $\ell > \ell'$ are GB classes.

$p_{gb,\ell}$ Probability that a GB-based class ℓ user will have a GB packet to transmit during a slot

$p_{gf,\ell}$ Probability that a class ℓ user will have a GF packet to transmit during a slot

σ_ℓ Parameter of the Bernoulli trial that an inactive class ℓ user generates a packet.

n_ℓ Number of class ℓ users in the system.

N Total number of users in the system.

N_{ps} Number of PS users in the system.

N_{gb} Number of GB users in the system.

M Total number of RUs available to serve all users.

M_{ps} Number of RUs allocated for dedicated service of periodic users.

M_{gb} Number of RUs allocated for dedicated service of grant-based users.

M_{gf} Number of RUs allocated for grant-free service for periodic and grant-based users.

b_ℓ Number of packets generated by inactive class ℓ users during a slot.

B_{gb} Number of new packets generated by inactive users of GB classes during a slot.

B_{gf} Number of new packets generated by inactive users of all classes during a slot.

As mentioned earlier, only inactive (idle) users are allowed to generate a packet. As defined above the probability that an inactive class ℓ user generates a new packet is modeled as a Bernoulli trial with probability σ_ℓ. We assume that,

$$\sigma_\ell > \sigma_{\ell+1} \qquad , \quad \ell = 1 \dots L-1 \tag{3.2}$$

We note that the higher packet generation parameter of a class, the more chance it will be classified as a PS class. If class $(\ell + 1)$ is classified as a PS class, then class ℓ will also be classified as a PS class. Thus, there is a dividing line between the classes, where classes $\ell \leq \ell'$ are classified as PS classes and classes $\ell > \ell'$ as GB classes. Since users in the system are classified as either PS-based or GB-based, the total number of users in the system, N, is given as,

$$N = N_{ps} + N_{gb} = \sum_{\ell=1}^{\ell\prime} n_\ell + \sum_{\ell=\ell\prime+1}^{L} n_\ell = \sum_{\ell=1}^{L} n_\ell \tag{3.3}$$

The total number of available RUs, M, are partitioned among the three different type of resource allocation schemes as,

$$M = M_{ps} + M_{gb} + M_{gf} \tag{3.4}$$

As explained above, each user alternates between active and inactive states. A user remains in the inactive state until it generates a packet and then it transits into the active state. In the active state, the user is serving a packet and the service time of a packet is $K + 1$ slots. Once a packet is served, the user transits to the inactive state. Note that in case of mixed numerology, where a GF repetition and a GB dedicated transmission do not have the same TTI duration (t_{TTI}), a slot is assumed to have the duration of the longer TTI between the two (NUM1: $t_{TTI} = 142.8\ \mu s$). Let τ_ℓ denote the number of slots that a class ℓ user spends in the inactive state during each visit to that state. The probability distribution of the number of slots that a class ℓ user spends in the inactive state is given by the geometric distribution,

$$\Pr(\tau_\ell = j) = \sigma_\ell(1 - \sigma_\ell)^{j-1} \qquad , \quad j = 1,2, \dots \tag{3.5}$$

Then the average number of slots that a user spends in each visit to inactive state is,

$$E[\tau_\ell] = \frac{1}{\sigma_\ell} \tag{3.6}$$

3.5.1 Blocking probability of GB transmissions

In case the BS cannot assign a dedicated RU to an active GB user in the next immediate slot following the $m = 3$ slots overhead period (mentioned in the previous section), the GB transmission of the user's packet gets blocked. In this subsection we obtain the blocking probability of a GB packet. As each user alternates between active and inactive states, we will refer to an active period followed by an inactive period as a cycle. The average duration of a cycle is given by $K + 1 + E[\tau_\ell]$. Since each user generates a new packet at the beginning of each cycle, probability that a GB-based class ℓ user will generate a new GB packet to transmit during a slot is given by,

$$\begin{aligned} p_{gb,\ell} &= \frac{1}{K + 1 + E[\tau_\ell]} \\ &= \frac{\sigma_\ell}{\sigma_\ell(K+1)+1} \end{aligned} \tag{3.7}$$

Probability distribution of the number of GB packets generated by inactive GB-based class ℓ users during a slot is given by,

$$\Pr\left(b_{gb,\ell} = j\right) = \binom{n_\ell}{j} \left(p_{gb,\ell}\right)^j \left(1 - p_{gb,\ell}\right)^{n_\ell - j} \tag{3.8}$$

where $b_{gb,\ell}$ denotes the number of GB packets generated by class ℓ users during a slot, given that $\ell > \ell'$. Therefore, the total number of new packets generated by inactive GB users, B_{gb}, during a slot is expressed as,

$$B_{gb} = \sum_{i=\ell'+1}^{L} b_{gb,i} \tag{3.9}$$

Finding the probability mass function (PMF) for B_{gb} can be complex, therefore, we find the probability generating function (PGF) instead,

$$
\begin{aligned}
B_{gb}(z) &= E\left[z^{B_{gb}}\right] = E\left[z^{\sum_{i=\ell\prime+1}^{L} b_{gb,i}}\right] \\
&= \prod_{i=\ell\prime+1}^{L} \left(p_{gb,i} z + 1 - p_{gb,i}\right)^{n_i}
\end{aligned}
\tag{3.10}
$$

The average number of GB packets generated during a slot is found as,

$$
\bar{B}_{gb} = \frac{d}{dz} B_{gb}(z)|_{z=1} \tag{3.11}
$$

Let us define the variable d as the number of blocked GB packets in a slot. Then, the average number of blocked packets in a slot $\bar{d}$ is found as follows,

$$
\begin{aligned}
\bar{d} &= \sum_{j=M_{gb}+1}^{N_{gb}} \left(j - M_{gb}\right) \times \Pr\left(B_{gb} = j\right) \\
&= \sum_{j=M_{gb}+1}^{N_{gb}} \mathrm{j} \Pr\left(B_{gb} = j\right) - M_{gb} \sum_{j=M_{gb}+1}^{N_{gb}} \Pr\left(B_{gb} = j\right) \\
&= \bar{B}_{gb} - \sum_{j=0}^{M_{gb}} j \Pr\left(B_{gb} = j\right) - M_{gb}\left[1 - \sum_{j=0}^{M_{gb}} \Pr\left(B_{gb} = j\right)\right] \\
&= \bar{B}_{gb} - M_{gb} + \left[\sum_{j=0}^{M_{gb}} \left(M_{gb} - j\right) \Pr\left(B_{gb} = j\right)\right] \\
&= \bar{B}_{gb} - M_{gb} + \bar{r}_{gb}
\end{aligned}
\tag{3.12}
$$

where $\bar{r}_{gb}$ is the average number of unused GB resource units in a slot. Hence, the blocking probability of a GB packet is given by,

$$
P_b = \bar{d} \big/ \bar{B}_{gb} \tag{3.13}
$$

To verify the accuracy of the blocking probability obtained in (3.13), we compare them with simulation results obtained through Matlab. In Fig. 3.4, the blocking probability values are obtained for the class parameters listed in Table 3.5, assuming M_{gb} values listed in Table 3.6. Each point in the horizontal axis corresponds to a different value of dividing class index. Results show that analysis is accurate as it matches the results obtained from simulation.

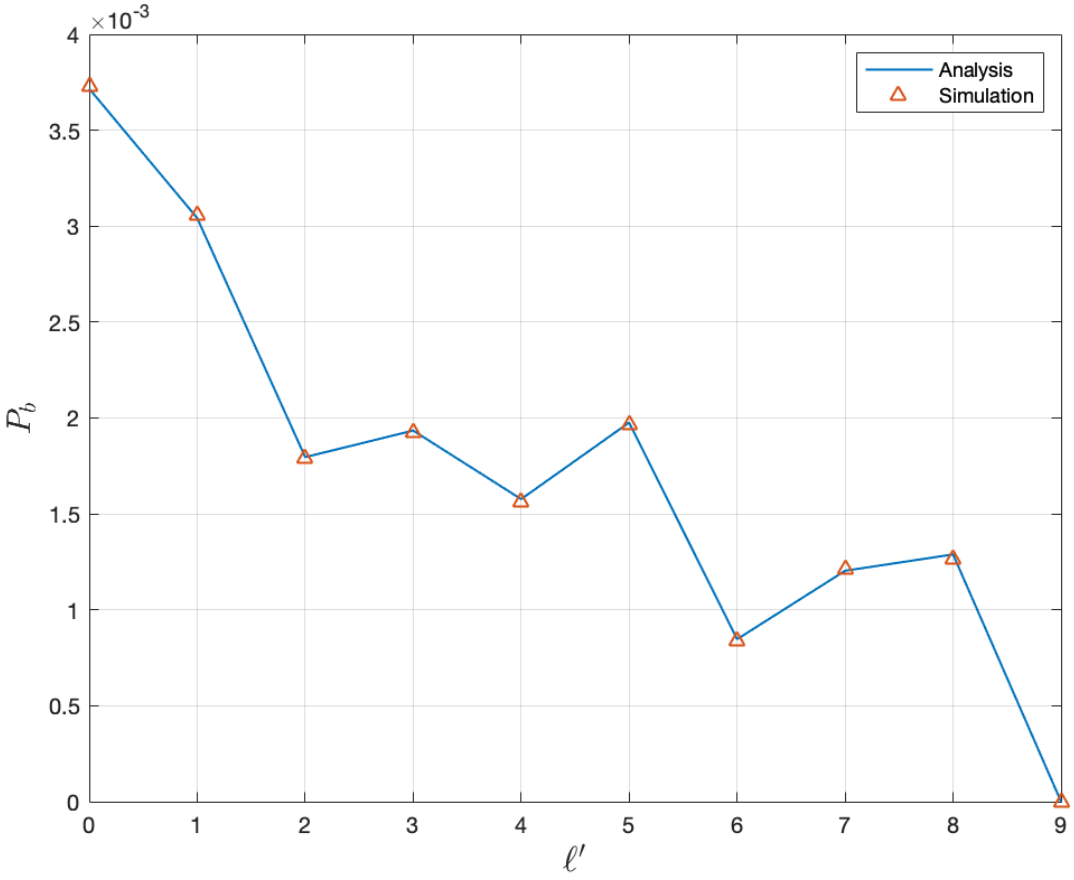

Fig. 3.4. Comparison between analytical and simulation results of the blocking probability.

Table 3.5. User class parameters.

$\boldsymbol{\ell}$	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
$\boldsymbol{\sigma_\ell}$	0.95	0.9	0.7	0.52	0.5	0.47	0.26	0.22	0.1
$\boldsymbol{n_\ell}$	5	6	4	5	7	4	8	11	20

Table 3.6. RUs allocated at each dividing class index point.

ℓ'	0	1	2	3	4	5	6	7	8	9
M_{gb}	10	9	8	7	6	5	5	4	3	0

3.5.2 Collision probability of GF repetitions

Probability that a class ℓ user will have a GF packet to transmit during a slot can be found in a similar way to (3.7). Since when a user becomes active it generates K repetitions to be transmitted as GF packets, probability that a class ℓ user will generate a GF packet during a slot is K times higher than the one in (3.7) and can be obtained as,

$$\begin{aligned} p_{gf,\ell} &= \frac{K}{K+1+E[\tau_\ell]} \\ &= \frac{K\sigma_\ell}{\sigma_\ell(K+1)+1} \end{aligned} \tag{3.14}$$

To find the number of new packets generated by users of all classes in a slot, we use the same method as in (3.9) and (3.10),

$$B_{gf} = \sum_{\ell=1}^{L} b_\ell \tag{3.15}$$

$$\begin{aligned} B_{gf}(z) &= E[z^{B_{gf}}] = E[z^{\sum_{\ell=1}^{L} b_\ell}] \\ &= \prod_{i=1}^{L} (p_{gf,i}z + 1 - p_{gf,i})^{n_i} \end{aligned} \tag{3.16}$$

To determine the probability that one of the repetitions of a packet will have a collision, P_{co}, let us define the random variable X as the number of users colliding with the user of interest. An active user is equally likely to choose one of the shared RUs for GF for transmission of each repetition of its packet. Assuming B_{gf} is known, the probability of x users colliding with the user of interest, given the number of available RUs for GF repetitions M_{gf}, is binomially distributed,

$$\Pr(X = x \mid B_{gf} = k, k > 0) = \binom{k-1}{x}\left(\frac{1}{M_{gf}}\right)^{x}\left(1 - \frac{1}{M_{gf}}\right)^{k-1-x} \tag{3.17}$$

where $\left(\frac{1}{M_{gf}}\right)$ represents the probability that an active user transmits a packet on the same RU used by the user of interest. From this, we can get the probability of a collision-free repetition for the user of interest as,

$$\Pr(X = 0 \mid B_{gf} = k, k > 0) = \left(\frac{M_{gf} - 1}{M_{gf}}\right)^{k-1} \tag{3.18}$$

$$\Pr(X = 0) = \sum_{k=1}^{N}\left(\frac{M_{gf} - 1}{M_{gf}}\right)^{k-1}\frac{\Pr(B_{gf} = k)}{1 - \Pr(B_{gf} = 0)} \tag{3.19}$$

Finally, the probability of collision can be expressed as,

$$P_{co} = 1 - \Pr(X = 0) = 1 - \sum_{k=1}^{N}\left(\frac{M_{gf} - 1}{M_{gf}}\right)^{k-1}\frac{\Pr(B_{gf} = k)}{1 - \Pr(B_{gf} = 0)} \tag{3.20}$$

A single GF repetition of a packet will be successful if it does not experience collision as well as channel degradation (ϵ). Grant-free transmissions of a packet will be successful if at least one transmission (out of the K repetitions) is successful. Let us define P_S probability that a grant-free transmission of a packet will be successful, then it is given by,

$$P_S = 1 - [1 - (1 - P_{co})(1 - \epsilon)]^{K} \tag{3.21}$$

where in the above $(1 - P_{co})(1 - \epsilon)$ corresponds to the probability that a single repetition will be successfully received. For the sake of verifying the analysis related to collisions through simulation, let us define the variable P_f as the probability of collision of all K repetitions without considering the packer error rate. This can be achieved by finding the complement of (3.21) and setting $\epsilon = 0$.

$$P_f = (P_{co})^K \tag{3.22}$$

For a single class network with $N = 30$ users and $K = 3$ repetitions, we find both analysis and simulation results of P_f under different σ values for $M_{gf} = 15$. The results plotted in Fig. 3.5 show the accuracy of the presented collision probability for GF repetitions.

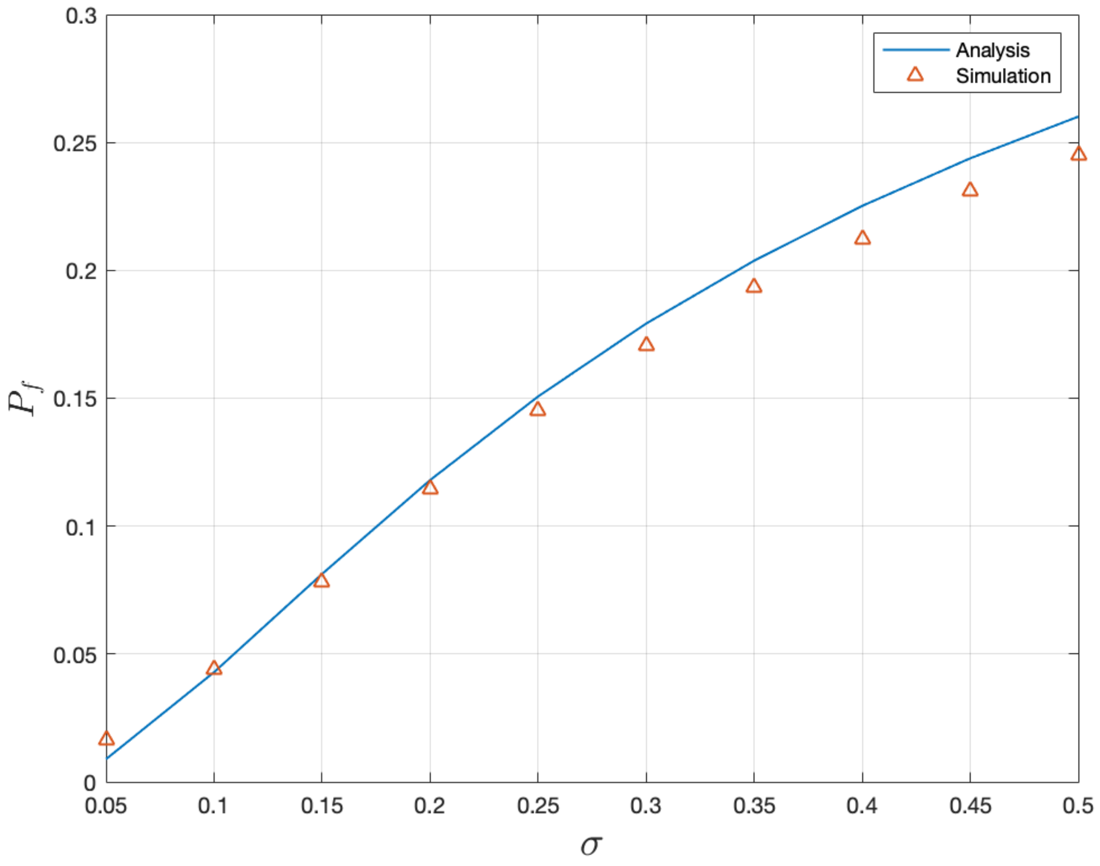

Fig. 3.5. Comparison between analytical and simulation results of the collision probability of all K repetitions.

3.5.3 Total Packet Loss Probabilities

Next, we determine the probability of packet loss for both PS-based and GB-based users. A packet will be lost if the dedicated transmission as well as all of the grant-free repetitions fail. For PS-based users, this probability is given by,

$$q_{ps} = \epsilon(1 - P_S) \tag{3.23}$$

where ϵ is the failure probability of the PS dedicated transmission and $(1 - P_S)$ is the failure probability of K GF repetitions. Since the PS transmission and GF repetitions are independent of each other, the probability of both events occurring is given by the product of their probabilities. In a similar manner, the probability of loss for GB users is expressed as,

$$q_{gb} = [1 - (1 - \epsilon)(1 - P_b)](1 - P_s) \tag{3.24}$$

where $(1 - \epsilon)(1 - P_b)$ is the probability of success for a GB transmission (i.e. the packet is not blocked and not lost due to channel degradation). Therefore, $1 - (1 - \epsilon)(1 - P_b)$ represents all the events leading to the loss of a GB dedicated-resource transmission of a packet and $(1 - P_s)$ denotes the probability that all K repetitions fail.

3.6 Conclusion

This chapter described the types of scheduling schemes considered in this thesis, as well as the context assumed for the network model. The latency of each of the three scheduling schemes was examined in order to check the validity of the proposed solution with regards to the latency requirements of URLLC. Lastly, we derived the equations leading to the probability of packet loss for both PS-based and GB-based users.

Chapter 4

System Optimization

4.1 Introduction

This chapter considers optimization of the system while meeting the reliability requirements of URLLC. Following the model in the previous chapter users are assumed to be divided into a number of classes according to their packet generation probabilities. We may choose either a heterogeneous or homogenous service strategy. In homogeneous strategy all user classes are served either as GF-based, PS-based or GB-based, these strategies are not efficient in resource utilization. The homogenous GF strategy requires huge amount of bandwidth to meet reliability requirements because of collisions. The homogenous PS strategy will not be efficient in resource utilization, since users in some classes generate packets very infrequently. On the other hand, the homogenous GB strategy may also not result in efficient resource utilization because of the substantial number of RUs needed to reduce the blocking probability in the presence of periodic traffic.

A heterogenous scheduling strategy may require less amount of resources compared to either of the homogenous scheduling strategies. We would like to determine the composition of the heterogeneous strategy, which user classes to be served as PS-based and which will be served as GB-based. Clearly, there will be a dividing line among the classes, all classes with packet generation probability higher than a threshold value will be served by the PS scheme, and those less than the threshold value by the GB scheme. We expect that this composition will depend on the objective of the optimization. System optimization is performed to either minimize the amount of resources needed or to maximize the traffic handling capacity. The main resource of the system is its bandwidth and in the following we will consider both optimization methods. In addition, the optimization of the system with respect to packet length has also been studied. All the optimization problems have been solved by using Fmincon optimization tool from Matlab.

4.2 Optimal resource allocation for bandwidth minimization

In this section, we optimize the system such that it results in the minimum amount of required bandwidth that satisfies the reliability requirements of URLLC. The objective of the optimization is to find the optimal resource allocation and dividing class index ℓ^* that results in the minimum amount of needed bandwidth W^*, given class parameters n_ℓ and σ_ℓ. Let us define W_{ps}, W_{gb} and W_{gf} as the amount of bandwidth (in Hz) allocated for PS, GB and GF scheduling, respectively. Their sum results in the total allocated bandwidth in the system,

$$W = W_{ps} + W_{gb} + W_{gf} \tag{4.1}$$

In order to show the effect of 5G NR's scalable numerology, we study system optimization for the two following cases,

- Single numerology: all scheduling schemes operate on NUM2.
- Mixed numerology: GB operating on NUM2, while PS and GF on NUM1.

From (2.3), let us define ϕ_{BW1} and ϕ_{BW2} as the bandwidth required (in Hz) for one RU operating on NUM1 and NUM2, respectively. The number of required resource blocks (ϕ_{RB}) to transmit an L_p-sized packet can be found from (2.1) as,

$$\phi_{RB} = \left\lceil \frac{L_p}{\phi_{sym} \times 12} \right\rceil \tag{4.2}$$

Assuming mixed numerology, the amount of allocated bandwidth for each scheduling scheme is obtained as,

$$W_{ps} \quad = M_{ps} \times \phi_{BW1} = M_{ps} \times 12 \times \Delta f \times \left\lceil \frac{L_p}{\phi_{sym1} \times 12} \right\rceil \tag{4.3}$$

$$W_{gb} \quad = M_{gb} \times \phi_{BW2} = M_{gb} \times 12 \times \Delta f \times \left\lceil \frac{L_p}{\phi_{sym2} \times 12} \right\rceil \tag{4.4}$$

$$W_{gf} \quad = M_{gf} \times \phi_{BW1} = M_{gf} \times 12 \times \Delta f \times \left\lceil \frac{L_p}{\phi_{sym1} \times 12} \right\rceil \tag{4.5}$$

Values of ϕ_{sym1} and ϕ_{sym2} need to be the same as the ones listed in Table 2.3 (i.e. $\phi_{sym1} = 2$ and $\phi_{sym2} = 7$). This is done to ensure meeting the latency requirement of URLLC, where ϕ_{sym1} and ϕ_{sym2} refer to the number of OFDM symbols per transmission for NUM1 and NUM2, respectively. The optimal resource allocation is obtained by finding the optimal set of values M_{ps}^{*}, M_{gb}^{*} and M_{gf}^{*}. Since M_{ps}^{*} depends on ℓ^{*}, it is not an optimization variable and it can be obtained from the feasibility condition of PS (3.1) as,

$$M_{ps}^{*} = \frac{\sum_{\ell=1}^{\ell^{*}} n_{\ell}}{K+1} \tag{4.6}$$

Note that the above bandwidth relations belong to the mixed numerology case. The same relations can be applied for the single numerology case with the difference of replacing ϕ_{BW1} with ϕ_{BW2} in (4.3) and (4.5). Given the list of assumed system constants in Table 4.1, we formulate the problem of minimizing the required bandwidth subject to reliability constraints of URLCC as follows,

$$\min_{M_{gb}, M_{gf}, \ell'} \quad W_{ps} + W_{gb} + W_{gf} \tag{4.7}$$

$$\text{s.t.} \quad q_{ps} \leq 10^{-5} \tag{4.8}$$

$$q_{gb} \leq 10^{-5} \tag{4.9}$$

Constraints (4.8) and (4.9) refer to the packet loss probabilities in (3.23) and (3.24), respectively. Thus, maximum packet loss probabilities of both types of users have been set to 10^{-5}. In the following, we will consider the bandwidth optimization under two different scenarios.

Table 4.1. System constants.

Constant	Value	Definition
K	3	Number of GF repetitions
k	256	Number of information bits to be transmitted
L_p	392	Packet length (in symbols)
β	2	Signal to Noise Ratio
ϵ	1×10^{-4}	Packet Error Rate
L	9	Number of user classes
ϕ_{sym1}	2	Number of OFDM symbols for NUM1
ϕ_{sym2}	7	Number of OFDM symbols for NUM2

Table 4.2. User class parameters.

$\boldsymbol{\ell}$	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
$\boldsymbol{\sigma_\ell}$	0.7	0.6	0.5	0.4	0.1	0.08	0.06	0.04	0.02
$\boldsymbol{n_\ell}$	5	6	4	5	7	4	8	11	20

4.2.1 Scenario A: Transmission over shared resources only

In scenario A, we study the case of relying solely on the GF repetitions over shared resources. Since we eliminate the GB and PS strategies in this case, the reliability constraint is given by,

$$1 - P_S = [1 - (1 - P_{co})(1 - \epsilon)]^K \leq 10^{-5} \tag{4.10}$$

where P_S is the probability of success of at least one GF repetition out of the total K repetitions and P_{co} is the probability of collision for one repetition. Since all users in this scenario transmit their packets through the GF scheme, the only optimization variable is M_{gf}. Formulation of the optimization problem for this scenario is given below,

$$\min_{M_{gf}} \quad W_{gf} \tag{4.11}$$

$$\text{s.t.} \quad 1 - P_S \leq 10^{-5} \tag{4.12}$$

where,

$$P_S = 1 - [1 - (1 - P_{co})(1 - \epsilon)]^K \tag{4.13a}$$

$$P_{co} = 1 - \sum_{k=1}^{N} \left(\frac{M_{gf} - 1}{M_{gf}}\right)^{k-1} \frac{\Pr(B_{gf} = k)}{1 - \Pr(B_{gf} = 0)} \tag{4.13b}$$

$$B_{gf}(z) = \prod_{i=1}^{L} (p_{gf,i} z + 1 - p_{gf,i})^{n_i} \tag{4.13c}$$

$$p_{gf,\ell} = \frac{K\sigma_\ell}{\sigma_\ell(K+1) + 1} \tag{4.13d}$$

4.2.2 Results

The plot in Fig. 4.1 shows the minimum required bandwidth to satisfy the reliability requirement for different values of K as a function of ϵ, assuming the system constants and class parameters listed in Table 4.1 and Table 4.2, respectively. From Table 4.1, number of user classes has been set to $L = 9$, and Table 4.2 gives number of users in each class and packet generation probability of each user class. Intuitively, the minimum bandwidth needed to satisfy the reliability constraint obtained in Fig. 4.1 increases as the packet error rate increases. However, since K is the exponent in the reliability constraint (4.10), W^* values become less susceptible to changes of ϵ as the value of K increases. Note that for the case $K = 1$, it is not possible to achieve the target reliability if $\epsilon > 10^{-5}$. This can be demonstrated by assuming the best-case scenario in where there are zero collisions ($P_{co} = 0$), in which case the constraint (4.10) becomes,

$$1 - P_S = [\epsilon]^K \leq 10^{-5} \tag{4.14}$$

Results also show that greater number of repetitions give better system performance in terms of required bandwidth, which is due to the increased probability of success P_S. However, the magnitude of the performance improvement becomes smaller as the value of K increases. The limitation here is the total latency of a packet which is calculated by multiplying the number of repetitions by the time needed for one transmission (Transmission Time Interval t_{TTI}). In subsection 3.4.3 we showed that the latency for one GF repetition is $t_{TTI} = 142.8\ \mu s$. Therefore, the maximum number of repetitions achievable without exceeding the $0.6\ ms$ latency is $K = 4$ repetitions.

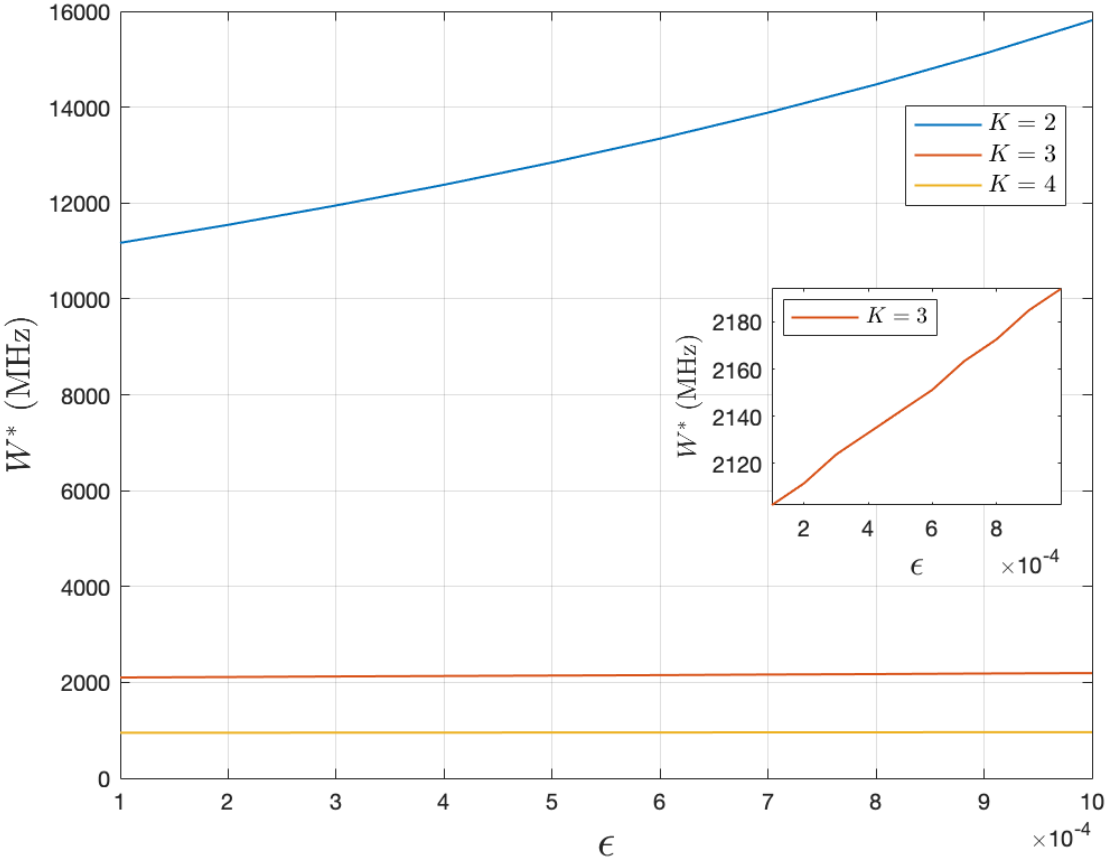

Fig. 4.1. Minimum required bandwidth for GF repetitions approach with different K values.

4.2.3 Scenario B: Transmission over both dedicated and shared resources

This scenario assumes packet transmissions using both dedicated and shared resources. Dedicated resources are accessed via GB or PS, depending on the user's class type which is determined by

ℓ'. On the other hand, shared resources are used to transmit the K repetitions through the GF scheme. As explained in the previous chapter, the dedicated transmission of the GB users will take place after the transmission of GF repetitions. The optimization variables in this scenario are: M_{gb}, M_{gf}, and ℓ'. As mentioned earlier, M_{ps} is not considered as an optimization variable because it depends on the value of the dividing class index as shown in (4.6). Formulation of the optimization problem for this scenario is given below,

$$\min_{M_{gb}, M_{gf}, \ell'} \quad W_{ps} + W_{gb} + W_{gf} \tag{4.15}$$

$$\text{s.t.} \quad \epsilon(1 - P_S) \leq 10^{-5} \tag{4.16}$$

$$[1 - (1 - \epsilon)(1 - P_b)](1 - P_s) \leq 10^{-5} \tag{4.17}$$

where,

$$P_S = 1 - [1 - (1 - P_{co})(1 - \epsilon)]^K \tag{4.18a}$$

$$P_{co} = 1 - \sum_{k=1}^{N} \left(\frac{M_{gf} - 1}{M_{gf}}\right)^{k-1} \frac{\Pr(B_{gf} = k)}{1 - \Pr(B_{gf} = 0)} \tag{4.18b}$$

$$B_{gf}(z) = \prod_{i=1}^{L} (p_{gf,i} z + 1 - p_{gf,i})^{n_i} \tag{4.18c}$$

$$p_{gf,\ell} = \frac{K\sigma_\ell}{\sigma_\ell(K + 1) + 1} \tag{4.18d}$$

$$P_b = {}^{\bar{d}}\!\big/_{\bar{B}_{gb}} \tag{4.18e}$$

$$\bar{d} = \sum_{j=M_{gb}+1}^{N_{gb}} \left(j - M_{gb}\right) \times \Pr\left(B_{gb} = j\right) \tag{4.18f}$$

$$B_{gb}(z) = \prod_{i=\ell'+1}^{L} (p_{gb,i} z + 1 - p_{gb,i})^{n_i} \tag{4.18g}$$

$$p_{gb,\ell} = \frac{\sigma_\ell}{\sigma_\ell(K + 1) + 1} \tag{4.18h}$$

4.2.4 Results

For this scenario, we find the minimum required bandwidth for the proposed mixed numerology case as well as the single numerology case. Results for both cases are shown in Fig. 4.2 where the minimum required bandwidth is plotted for different values of the dividing class index ℓ'. Each point in this figure is obtained by solving the optimization problem in (4.15) - (4.18) by treating ℓ' as a constant instead of an optimization variable. Then the value of ℓ' leading to the lowest W^* will evidently be the optimal dividing class index ℓ^*. Note that $\ell' = 0$ and $\ell' = 9$ correspond to homogenous GB and PS scheduling strategies discussed above, respectively.

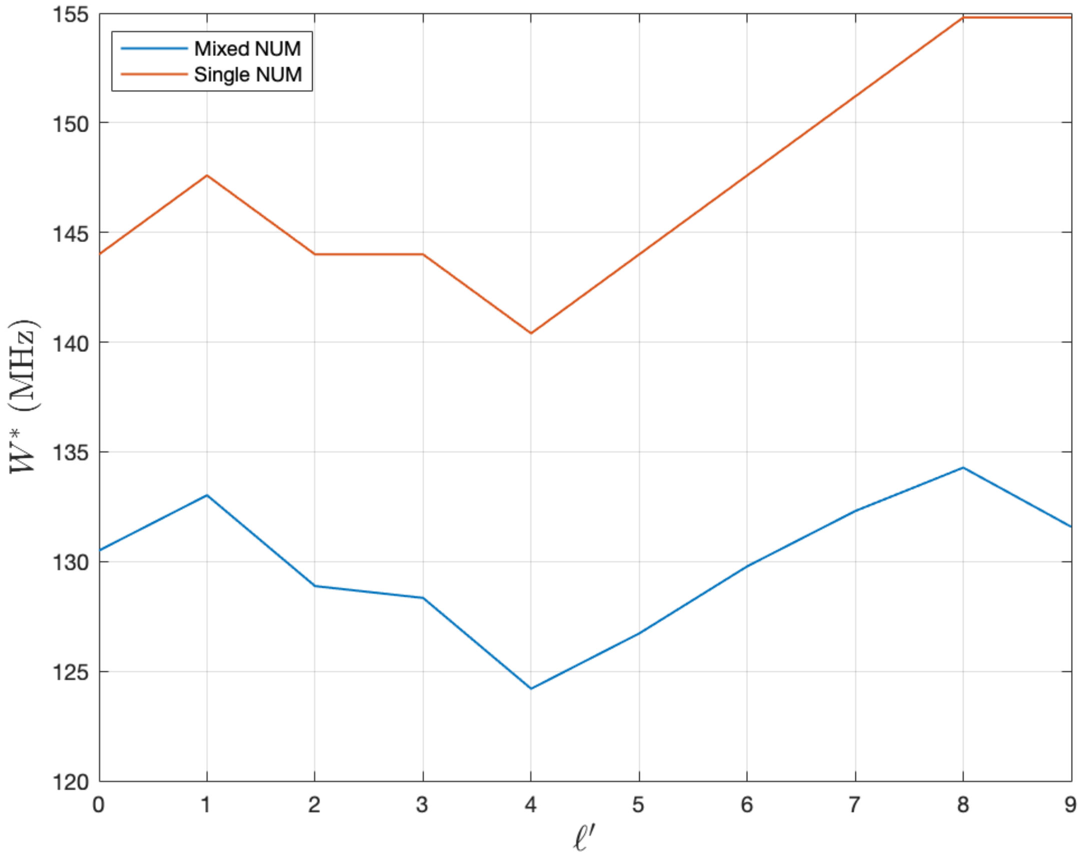

Fig. 4.2. Minimum required bandwidth as a function of the dividing class indices for mixed and single numerology cases in scenario B.

As expected, the optimal dividing class index value obtained is the one that assigns classes with high σ_ℓ values as PS-based classes while classes with low σ_ℓ values as GB-based classes. It is found that the optimal dividing index for scenario B is $\ell^* = 4$ for both numerology cases thus classes $1 \leq \ell \leq \ell^*$ will be served by PS scheduling scheme and classes $\ell > \ell^*$ by GB scheme. The optimal resource allocation obtained is $M_{ps}^* = 5$, $M_{gb}^* = 9$ and $M_{gf}^* = 25$. This results in a minimum required bandwidth of $W^* = 124.2$ MHz for the mixed numerology case, while for the single numerology case it is found to be $W^* = 140.4$ MHz. It is clear from the figure that the mixed numerology case is the superior one, this is owed to relying on NUM1 for PS and GF schemes, which was proved (in section 2.3) to require less bandwidth at the expense of a slightly higher latency. Therefore, in the remainder of this chapter we will assume a mixed numerology setting for all the remaining scenarios, unless indicated otherwise.

Table 4.3 shows the optimal allocation of resource units for each dividing class index value (ℓ'). From the table we can see that when $\ell' = 0$, $M_{gb} = 15$ and $M_{ps} = 0$ (because all users are considered as GB users), hence the total required RUs (including M_{gf}) is 40. On the other hand, when $\ell' = 1$, $M_{gb} = 14$ and $M_{ps} = 2$ (because there are 5 PS users which will need 2 RUs), so the total required RUs is 41. This explains why the bandwidth required for the point $\ell' = 0$ is smaller than the point $\ell' = 1$. Similarly, for the points $\ell' = 8$ and $\ell' = 9$, we can see that the total required RUs is 43 for both points. Recall that GB operates on NUM2 while GF operates on NUM1. In Fig. 2.1 we showed that an RU using NUM2 requires more bandwidth than a NUM1 RU. So, although points $\ell' = 8$ and $\ell' = 9$ required the same number of resource units (43 RUs), point $\ell' = 8$ has 5 RUs operating on NUM2. This explains why the bandwidth required for the point $\ell' = 9$ is smaller than the point $\ell' = 8$ in the mixed numerology case.

Table 4.3. Optimal resource allocation for scenario B (single and mixed numerology).

$\boldsymbol{\ell'}$	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
$\boldsymbol{M_{ps}}$	0	2	3	4	5	7	8	10	13	18
$\boldsymbol{M_{gb}}$	15	14	12	11	9	8	8	7	5	0
$\boldsymbol{M_{gf}}$	25	25	25	25	25	25	25	25	25	25

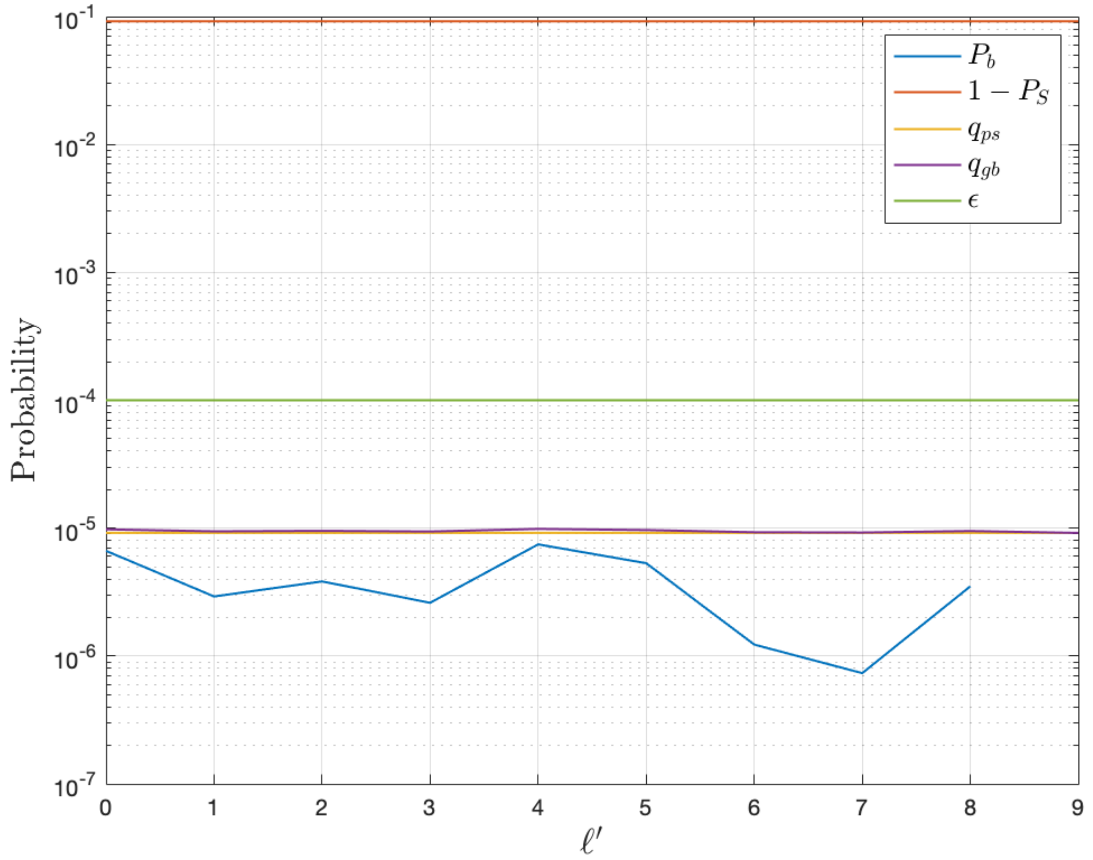

Fig. 4.3. Total packet loss probabilities for PS and GB users (q_{ps}, q_{gb}), packet blocking probability of GB users (P_b), packet loss probability of GF repetitions (1-P_S) and packet error rate (ϵ) as a function of the dividing class index in scenario B.

In Fig. 4.3, we plot values of the different probability components that form the optimization problem's constraints in (4.16) and (4.17). Since the objective is to minimize the bandwidth, the optimization program finds the minimum number of required resource units to satisfy the constraints (reliability requirement). As a result, the values of packet loss probabilities q_{ps} and q_{gb} will always be close to the reliability requirement (i.e. 10^{-5}). The probability of failure in the GF repetitions is constant as amount of GF traffic is independent of the assignment of user classes to the PS and GB schemes. We can also notice that the probability of failure in the GF repetition ($1 - P_S$) is relatively high (due to collisions), however this is sufficient to bring total packet loss probability below the threshold of 10^{-5}.

In Fig. 4.4, we study the effect of channel conditions ϵ on the amount of required bandwidth W^* as well as compare results from scenarios A and B, by applying the same constants and class parameters for the following cases,

- GF repetitions only ($K = 4$) from scenario A.
- Optimal heterogenous mixed numerology ($\ell^* = 4$) from scenario B.
- Homogenous GB-based strategy ($\ell' = 0$) from scenario B.
- Homogenous PS-based strategy ($\ell' = 9$) from scenario B.

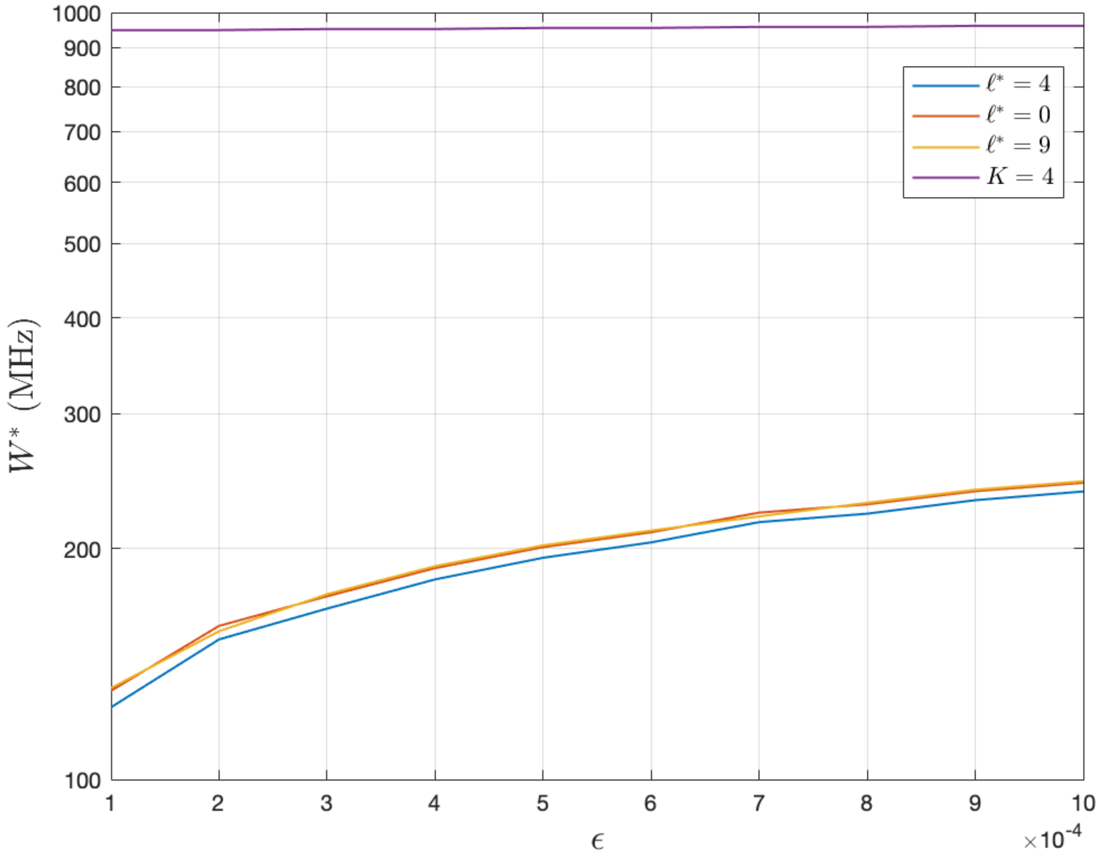

Fig. 4.4. Minimum required bandwidth under variable channel conditions for scenarios A and B.

As expected, the bandwidth requirement increases as the channel conditions get worse with increasing packet error rate. It is also shown that the case of our optimized mixed-numerology model utilizing all three schemes (PS, GB and GF) outperforms all of the other cases in terms of resource efficiency. The homogeneous GB-based and PS-based strategies contend for the second place as their resulting W^* values are very close with no significant superiority of one strategy over the other. The worst case was proved to be the GF repetition scheme, which verifies that GF repetitions should be considered only as supplementary to the dedicated-resource transmissions, rather than an independent transmission scheme.

4.3 Optimal packet length for bandwidth minimization

4.3.1 Problem Formulation

In this section, we optimize the system such that it determines the optimal packet length L_p^* leading to the minimum required bandwidth while meeting the reliability requirements of URLLC. Packet length has a trade-off nature when it comes to the required bandwidth. From (2.9), packet error rate is shown in (4.19) for convenience, where k is the number of information bits in the packet and L_p is the packet length. For a constant k, increasing the packet length is convenient for the packet error rate (ϵ), as it makes the probability of error smaller and hence requiring fewer resource units to satisfy the reliability requirement.

$$\epsilon \approx Q\left(\frac{L_p C - k + (log\, L_p)/2}{\sqrt{L_p V}}\right) \tag{4.19}$$

However, a longer packet length also means that more resource blocks are required per resource unit, resulting in an increase in the required bandwidth.

$$\phi_{RB} = \left\lceil \frac{L_p}{\phi_{sym} \times 12} \right\rceil \tag{4.20}$$

The objective is to find the perfect balance of this trade-off that would result in the least amount of bandwidth needed, assuming the blocking probability (3.13). The optimization variables in this section are: M_{gb}, M_{gf}, ℓ', and L_p. The optimization problem is formulated as follows,

$$\min_{M_{gb}, M_{gf}, \ell', L_p} \quad W_{ps} + W_{gb} + W_{gf} \tag{4.21}$$

$$\text{s.t.} \quad \epsilon(1 - P_S) \le 10^{-5} \tag{4.22}$$

$$[1 - (1 - \epsilon)(1 - P_b)](1 - P_s) \le 10^{-5} \tag{4.23}$$

where probability components in (4.22) and (4.23) are found from the set of relations in (4.18) and ϵ is given by (4.19). However, due to the added complexity of this optimization problem, we had to simplify (4.19) in order to make the process of solving it easier for Matlab. The Q-function in (4.19) is related to the complementary error function '*erfc*' as,

$$Q(x) = \frac{1}{2} erfc\left(\frac{x}{\sqrt{2}}\right) \tag{4.24}$$

For calculating the complementary error function, we used the pure exponential asymptotic approximation [23],

$$erfc(y) \approx \frac{1}{6} e^{-y^2} + \frac{1}{2} e^{-\frac{4}{3}y^2} \qquad , \quad y > 0 \tag{4.25}$$

Therefore, the packet error rate relation in (4.19) can be expressed as,

$$\epsilon \approx \frac{1}{2}\left(\frac{1}{6} e^{-y^2} + \frac{1}{2} e^{-\frac{4}{3}y^2}\right) \qquad , \quad y > 0 \tag{4.26}$$

where $y = \frac{1}{\sqrt{2}}\left(\frac{L_p C - k + (\log L_p)/2}{\sqrt{L_p V}}\right)$.

4.3.2 Results

The optimization problem is solved for the class parameters listed in Table 4.2 and system constants in Table 4.4. The optimal packet length is found to be the range $L_p^* = [412 - 420]$ symbols with optimal dividing class index $\ell^* = 4$, resulting in a minimum bandwidth of $W^* = 51.84$ MHz.

Table 4.4. System constants.

Constant	Value	Definition
K	3	Number of GF repetitions
k	256	Number of information bits to be transmitted
β	2	Signal to Noise Ratio
L	9	Number of user classes
ϕ_{sym1}	2	Number of OFDM symbols for NUM1
ϕ_{sym2}	7	Number of OFDM symbols for NUM2

In Fig. 4.5, we plot the minimum required bandwidth as a function of the packet length L_p for the optimal dividing class index $\ell^* = 4$. The plot has a step nature indicating equal amount of bandwidth required for multiple adjacent packet lengths. This is due to the ceiling function in (4.20), forcing ϕ_{RB} to have only integer values. Since multiple packet lengths result in the same value of ϕ_{RB}, the optimal packet length is found to be more than one value (i.e. $L_p^* = [412 - 420]$). Moreover, the optimal resource allocation for PS, GB, and GF is $M_{ps}^* = 5$, $M_{gb}^* = 9$ and $M_{gf}^* = 1$, respectively.

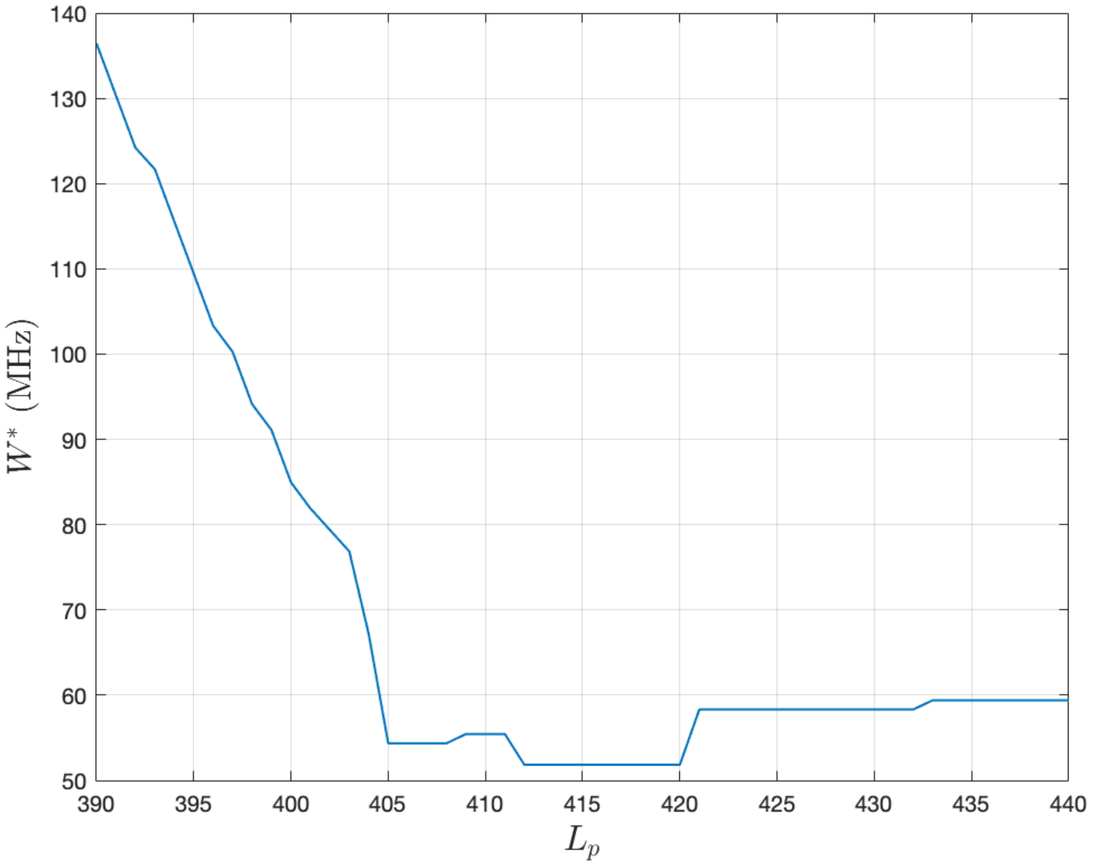

Fig. 4.5. Minimum required bandwidth as a function of packet length for $\ell^* = 4$.

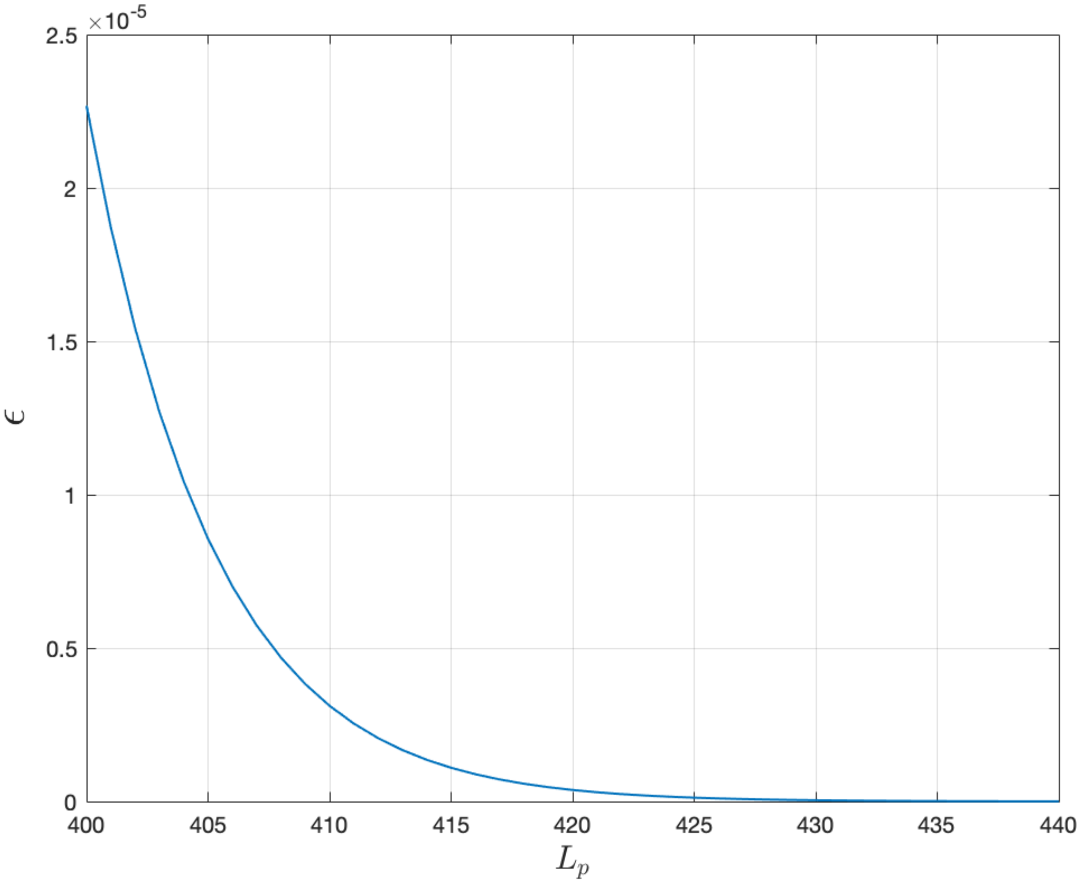

Fig. 4.6. Packet error rate as a function of packet length for the constants assumed in Table 4.1.

In Fig. 4.6, we plot the resulting packet error rate for the values of packet length considered in Fig. 4.5, assuming the system constants listed in Table 4.1. For the optimal packet length values L_p^* listed in Table 4.5 at each dividing class index value ℓ', we plot the minimum bandwidth required for different values of dividing class index ℓ' in Fig. 4.7. This was obtained by setting the dividing class index as a constant and solving the optimization problem for all possible values, as done in previous cases. We can see from the figure that the optimal dividing class index is $\ell^* = 4$, resulting in a minimum required bandwidth of $W^* = 51.84$ MHz as it was also shown previously in Fig. 4.5.

Table 4.5. Optimal packet length at each dividing class index value.

ℓ'	0	1	2	3	4	5	6	7	8	9
L_p^*	$[410 - 420]$	$[406 - 408]$	$[407 - 408]$	$[406 - 408]$	$[412 - 420]$	$[409 - 420]$	$[405 - 408]$	$[405 - 408]$	$[407 - 408]$	$[405 - 408]$

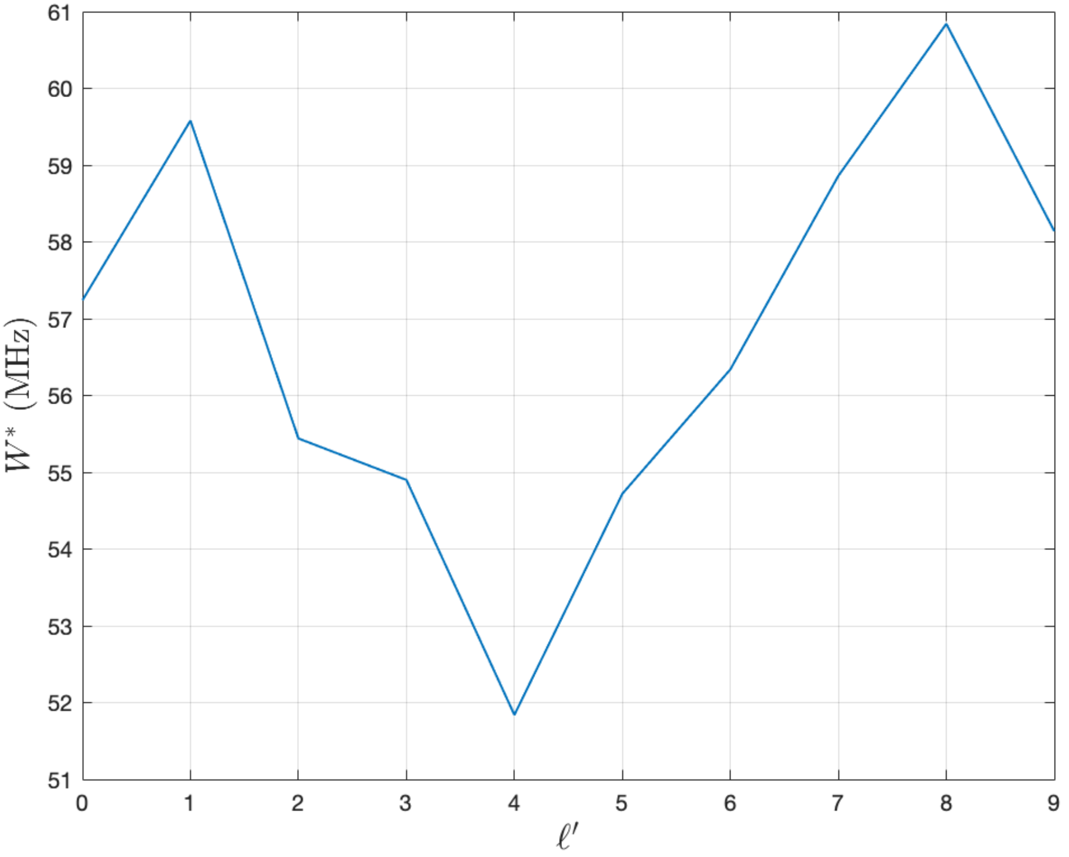

Fig. 4.7. Bandwidth required for different values of dividing class index for $L_p^* = 412$.

4.4 Optimal modulation and coding scheme for bandwidth minimization

4.4.1 Problem Formulation

In this section, we will be studying the effect of the Modulation and Coding Scheme (MCS) on the bandwidth requirement. The optimization problem posed in the previous section assumed that the packet length can take any value, hence giving total flexibility to the coding rate. However, for this problem we will strictly consider the MCS cases suggested by the standards of 3GPP in [24] and listed in Table 4.6. These cases are characterized based on their spectral efficiency η, which evaluates each MCS according to the number of information (useful) bits per modulation symbol. Since the code rate R is the ratio of information bits to the number of transmitted symbols ($R = {}^{k}/_{L_p}$), the spectral efficiency is given by,

$$\eta = R \times Q_m \tag{4.27}$$

where the modulation order Q_m is the number of bits transmitted per symbol. The optimization variables in this section are: M_{gb}, M_{gf}, ℓ', and I_{MCS}. Note that I_{MCS} determines the value of the packet length L_p through the code rate R given for each MCS. The optimization problem is formulated as follows,

$$\min_{M_{gb},M_{gf},\ell',I_{MCS}} \quad W_{ps} + W_{gb} + W_{gf} \tag{4.28}$$

$$\text{s.t.} \quad \epsilon(1 - P_S) \leq 10^{-5} \tag{4.29}$$

$$[1 - (1 - \epsilon)(1 - P_b)](1 - P_s) \leq 10^{-5} \tag{4.30}$$

where probability components in (4.29) and (4.30) are found from the set of relations in (4.18) and ϵ is given by (4.19). The objective of this optimization problem is to find the minimum amount of bandwidth required for different MCS cases which allows us to determine the optimal MCS for this URLLC model, as well as have insight on whether it is better to rely on a conservative MCS (low η) or an efficient MCS (high η).

Table 4.6. MCS table [24].

MCS Index ($\boldsymbol{I_{MCS}}$)	**Modulation Order** ($\boldsymbol{Q_m}$)	**Code Rate** ($\boldsymbol{R}$)	**Packet Length** ($\boldsymbol{L_p}$) (for $\boldsymbol{k = 256}$)	**Packet Error Rate** ($\boldsymbol{\epsilon}$) (for $\boldsymbol{k = 256}$ and $\boldsymbol{\beta = 1.26}$)	**Spectral Efficiency** ($\boldsymbol{\eta}$)
0	2	30/1024	8739	0	0.0586
1	2	40/1024	6554	0	0.0781
2	2	50/1024	5243	0	0.0977
3	2	64/1024	4096	0	0.1250
4	2	78/1024	3361	0	0.1523
5	2	99/1024	2648	3.6507e-244	0.1934
6	2	120/1024	2185	6.2464e-186	0.2344
7	2	157/1024	1670	2.0010e-122	0.3066
8	2	193/1024	1359	3.3565e-85	0.3770
9	2	251/1024	1045	1.3433e-49	0.4902
10	2	308/1024	852	1.5088e-29	0.6016
11	2	379/1024	692	4.0992e-15	0.7402
12	2	449/1024	584	2.8923e-07	0.8770
13	2	526/1024	499	0.0072	1.0273
14	4	340/1024	772	5.6483e-22	1.3281
15	4	378/1024	694	2.9057e-15	1.4766
16	4	434/1024	605	1.4332e-08	1.6953
17	4	490/1024	535	1.6729e-04	1.9141
18	6	438/1024	599	3.3236e-08	2.5664
19	6	466/1024	563	5.4859e-06	2.7305
20	6	517/1024	508	0.0032	3.0293

4.4.2 Results

In Fig. 4.8, we plot the minimum required bandwidth W^* at the optimal point ℓ^* for each of the MCS cases listed in Table 4.6, assuming the system constants and class parameters listed in Table 4.4 and Table 4.2, respectively. An extremely conservative MCS such as [$I_{MCS} = 0$ - $I_{MCS} = 2$] leads to having a longer packet to transmit a fixed number of information bits k. This provides extra redundancy which results in a smaller packet error rate ϵ. However, as it was mentioned in the previous section, this also leads to an increase in the number of RBs needed for one transmission. This justifies the high amount of bandwidth required for the cases [$I_{MCS} = 0$ - $I_{MCS} = 2$]. On the other hand, a highly efficient MCS with a very short packet length (e.g. $I_{MCS} = 13$ and $I_{MCS} = 20$) results in a high value of ϵ, which would require a large number of RUs allocated to the GF shared pool in order to compensate the low probability of decoding a collision-free packet. It is found that the optimal MCS is $I_{MCS}^{*} = 12$ which is considered to have moderate efficiency compared to the other considered MCS cases, while also having a packet length that results in a sufficiently low packet error rate ϵ, offering the optimal balance in this trade-off. The MCS value of $I_{MCS}^{*} = 12$ corresponds to a packet length of $L_p = 584$ symbols for $k = 256$ information bits. The optimal values of the optimization variables are: $\ell^{*} = 4$ resulting in $M_{ps}^{*} = 5$, $M_{gb}^{*} = 9$ and $M_{gf}^{*} = 1$. These set of values result in a minimum required bandwidth of $W^{*} = 72.36$ MHz.

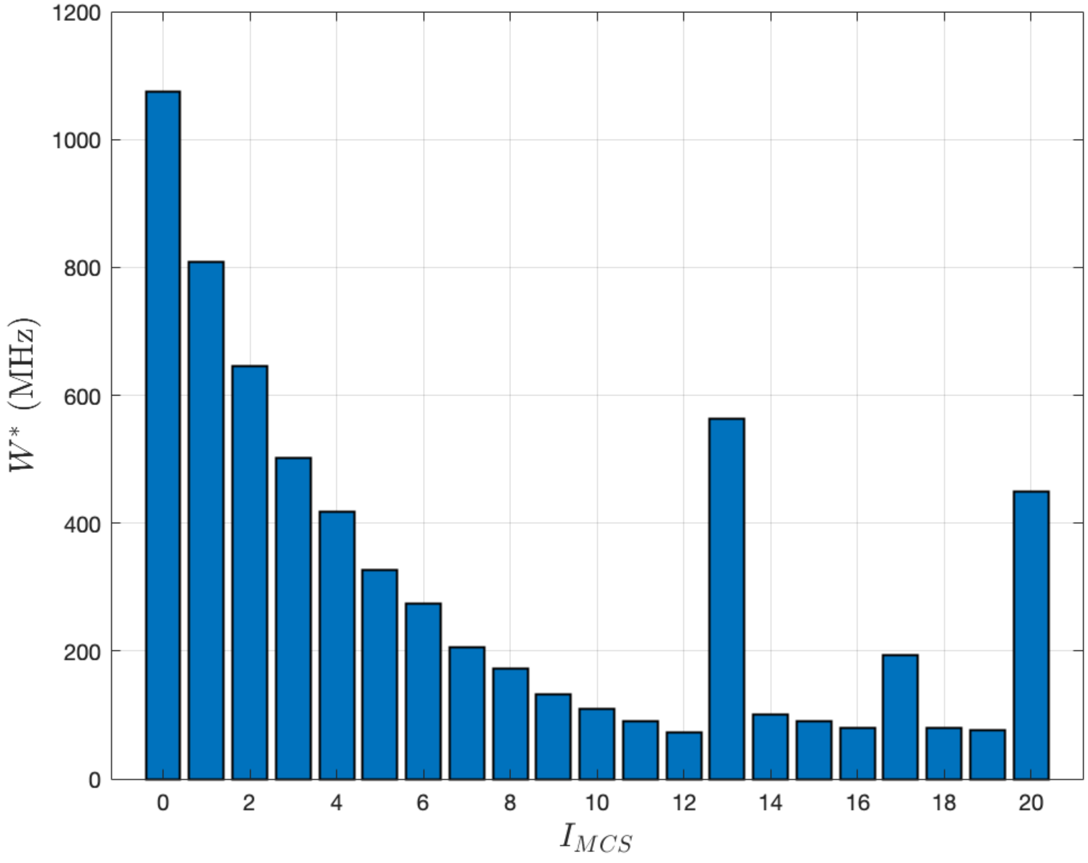

Fig. 4.8. Minimum required bandwidth for different MCS cases.

4.5 System optimization for maximum traffic capacity

4.5.1 Scenario A: Single class of users

In this scenario, we optimize a single-class system to maximize the traffic capacity for a given amount of resources while satisfying the reliability requirements of URLLC. To formulate the problem, let us first define the traffic T as,

$$T = \sigma n \tag{4.31}$$

where σ is the probability an inactive user will generate a packet, while n is the number of users in the network. For the consistency of resource units' dimensions, we assume a single numerology model where all schemes operate on NUM2. The optimization variables for the case of PS strategy are: M_{ps}, M_{gf}, and n, while for the GB strategy they are: M_{ps}, M_{gf}, and n. Note that optimal value M_{ps}^{*} in this scenario is calculated as,

$$M_{ps}^{*} = \frac{n^{*}}{K+1} \tag{4.32}$$

For a given number of available resource units M and Bernoulli parameter σ, the optimization problem can be formulated as,

- In case users are assumed to be PS-based

$$\max_{M_{ps}, M_{gf}, n} \quad T \tag{4.33}$$

$$\text{s.t.} \quad \epsilon(1 - P_S) \leq 10^{-5} \tag{4.34}$$

$$M_{ps} + M_{gf} = M \tag{4.35}$$

- In case users are assumed to be GB-based

$$\max_{M_{gb},M_{gf},n} \quad T \tag{4.36}$$

$$\text{s.t.} \quad [1-(1-\epsilon)(1-P_b)](1-P_s) \leq 10^{-5} \tag{4.37}$$

$$M_{gb}+M_{gf}=M \tag{4.38}$$

where probability components in the constraints (4.34) and (4.37) can be obtained from the set of relations in (4.18) and ϵ is given by (4.19).

Table 4.7. System constants for scenario A.

Constant	Value	Definition
K	3	Number of GF repetitions
k	256	Number of information bits to be transmitted
L_p	392	Packet length (in symbols)
β	2	Signal to Noise Ratio
ϵ	1×10^{-4}	Packet Error Rate
L	1	Number of user classes
ϕ_{sym2}	7	Number of OFDM symbols for NUM2
M	13	Number of available RUs

4.5.2 Results

We optimize a single numerology system operating on NUM2 with a channel bandwidth of 50 MHz, which results in $\psi_{RB}=65$ resource blocks, as it was shown in Table 2.4. Assuming the system constants in Table 4.11, a packet length $L_p=392$ requires $\phi_{RB}=5$ resource blocks per transmission from (4.20). Thus, the total number of available resource units per slot is $M=\frac{\psi_{RB}}{\phi_{RB}}=13$. We plot the maximum amount of supported traffic in Fig. 4.9 for the values of σ listed in Table 4.8, once assuming a PS class and another time assuming a GB class.

Table 4.8. Considered σ values in scenario A.

σ	0.01	0.03	0.05	0.07	0.09	0.11	0.13	0.15	0.17	0.19

Results show that for cases where $\sigma \leq 0.05$, the optimal choice of scheduling for users in the network is the GB strategy, while the PS strategy has shown to be superior in cases of higher traffic ($\sigma > 0.05$). This can be justified through the fact that the PS strategy is not affected by the values of σ, unlike the GB strategy where packets might get blocked more often in higher traffic conditions. The maximum number of supported users and optimal resource allocation at each σ values are listed in Table 4.9 and Table 4.10, for PS and GB strategies respectively.

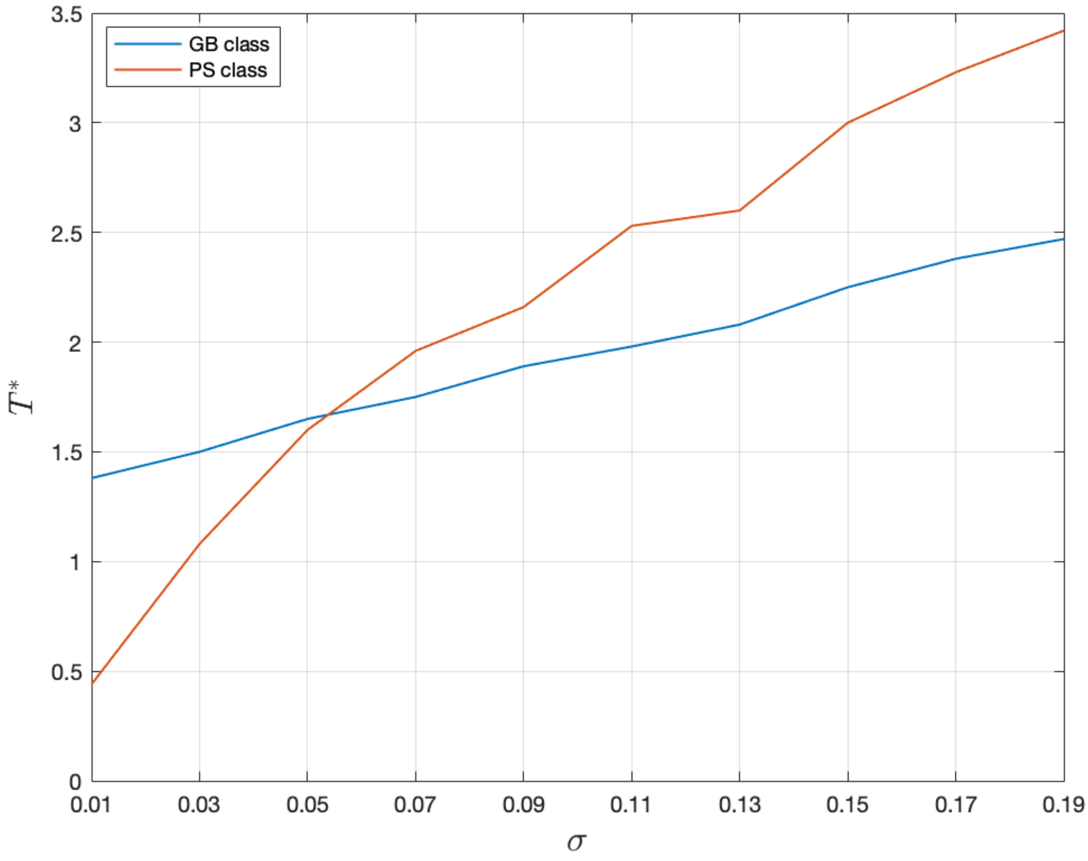

Fig. 4.9. Maximum traffic capacity as function of σ for scenario A.

Table 4.9. Optimal values in scenario A assuming PS strategy.

$\boldsymbol{\sigma}$	**0.01**	**0.03**	**0.05**	**0.07**	**0.09**	**0.11**	**0.13**	**0.15**	**0.17**	**0.19**
$\boldsymbol{n}^*$	44	36	32	28	24	23	20	20	19	18
$\boldsymbol{M_{ps}^*}$	11	9	8	7	6	6	6	5	5	5
$\boldsymbol{M_{gf}^*}$	2	4	5	6	7	7	7	8	8	8

Table 4.10. Optimal values in scenario A assuming GB strategy.

$\boldsymbol{\sigma}$	**0.01**	**0.03**	**0.05**	**0.07**	**0.09**	**0.11**	**0.13**	**0.15**	**0.17**	**0.19**
$\boldsymbol{n}^*$	138	50	33	25	21	18	16	15	14	13
$\boldsymbol{M_{gb}^*}$	8	8	7	7	7	7	7	7	7	7
$\boldsymbol{M_{gf}^*}$	5	5	6	6	6	6	6	6	6	6

4.5.3 Scenario B: Multiple classes of users

In this scenario, we optimize a multi-class system to maximize the traffic capacity for a given amount of resources while satisfying the reliability requirements of URLLC. Since there is more than one class of users, the traffic T is given by,

$$T = \sum_{\ell=1}^{L} \sigma_\ell n_\ell \tag{4.39}$$

where σ_ℓ is the probability an inactive class ℓ user will generate a packet, while n_ℓ is the number of users in that class. Similar to scenario A above, we assume that all schemes operate on NUM2. The optimization variables for this scenario are: M_{gb}, M_{gf}, ℓ, and n_ℓ. For a given number of available resource units M and Bernoulli parameter σ_ℓ, the optimization problem can be formulated as,

$$\max_{M_{gb},M_{gf},\ell',n_\ell} \quad T \tag{4.40}$$

$$\text{s.t.} \quad \epsilon(1-P_S) \leq 10^{-5} \tag{4.41}$$

$$[1-(1-\epsilon)(1-P_b)](1-P_s)) \leq 10^{-5} \tag{4.42}$$

$$M_{ps} + M_{gb} + M_{gf} = M \tag{4.43}$$

where probability components in the constraints (4.41) and (4.42) can be obtained from the set of relations in (4.18) and ϵ is given by (4.19).

Table 4.11. System constants for scenario B.

Constant	Value	Definition
K	3	Number of GF repetitions
k	256	Number of information bits to be transmitted
L_p	392	Packet length (in symbols)
β	2	Signal to Noise Ratio
ϵ	1×10^{-4}	Packet Error Rate
L	4	Number of user classes
ϕ_{sym2}	7	Number of OFDM symbols for NUM2
M	13	Number of available RUs

4.5.4 Results

Similar to scenario A, we optimize a single numerology system operating on NUM2 with a channel bandwidth of 50 MHz, resulting in the total number of available resource units per slot $M = 13$. For the system constants listed in Table 4.11 and class parameters listed in Table 4.12, we plot the maximum amount of supported traffic in Fig. 4.10 for different diving class indices, where the optimal results are found for each value of ℓ' separately.

Results show that the maximum achievable traffic capacity is $T^* = 3.45$ for optimal dividing class index $\ell^* = 4$. This means that all users are assumed to be PS-based, which proves that the PS strategy is the optimal one for high amount of traffics since it is not affected by the values of σ, unlike the blocking-prone strategy of GB scheduling. As a result, most of the users are assigned to the class with the highest σ value (class 1). The optimal distribution of users is: $n_1^* = 15, n_2^* = 2, n_3^* = 1$ and $n_4^* = 1$, while the 13 available resource units are distributed as: $M_{ps}^* = 5$, $M_{gb}^* = 0$ and $M_{gf}^* = 8$. We conclude that in the context of achieving a system that can support the maximum amount of traffic for a limited amount of assigned bandwidth, the homogeneous PS-based strategy is the optimal choice for this system.

Table 4.12. User class parameters.

ℓ	1	2	3	4
σ_ℓ	0.2	0.15	0.1	0.05

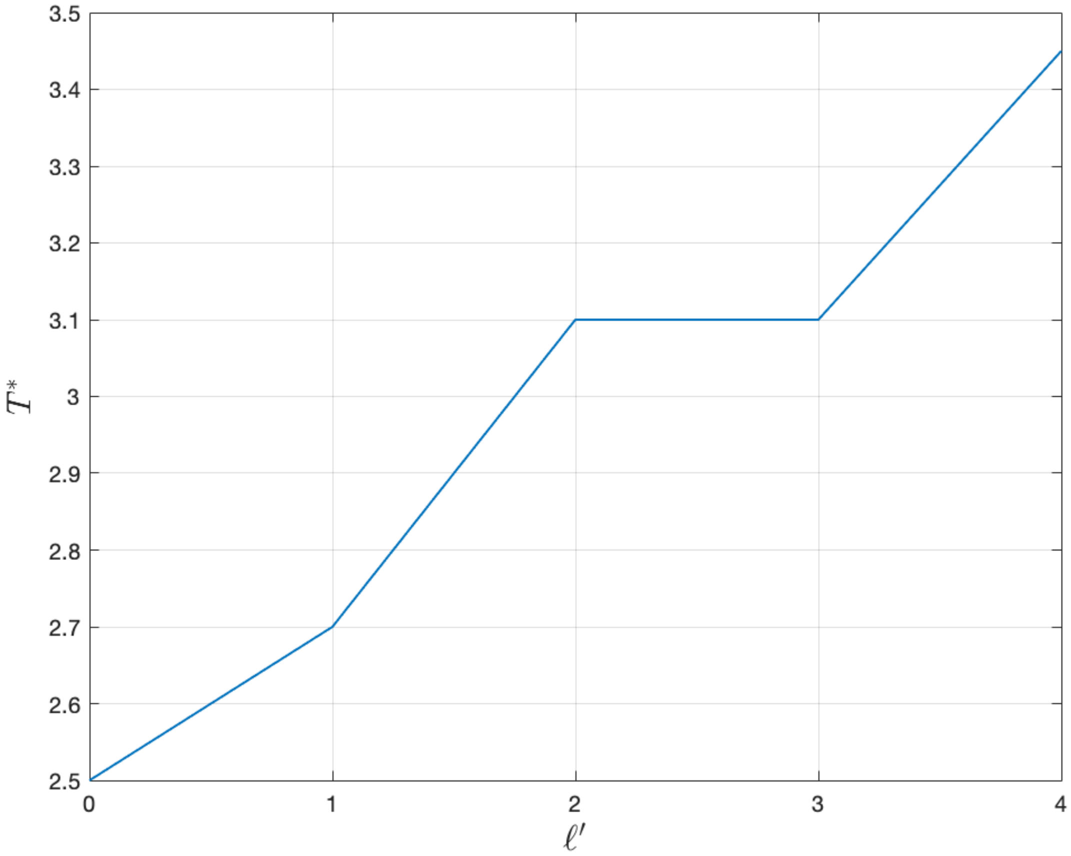

Fig. 4.10. Maximum traffic capacity for different diving class indices for scenario B.

4.6 Conclusion

In this chapter, we optimized the proposed system for either minimizing the required bandwidth or maximizing the traffic capacity of the system. In order to minimize the bandwidth requirement, we needed to assume the class parameters of the system, namely, the number of users in each class as well as the probabilities of packet generation of inactive users. On the other hand, for maximizing the traffic capacity, we needed to assume the amount of bandwidth available in the system as well as the probabilities of packet generation for each class. Different optimization problems were implemented for different cases or scenarios. However, for all the cases, we needed to find the optimal resource allocation for each considered scheduling scheme. Other optimal parameters were obtained that are specific for each of the discussed cases.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

The problem of URLLC is the most demanding prospective service of 5G NR. It is clear that existing communication systems (e.g. 4G LTE) are not capable of achieving the objectives and requirements of URLLC applications. Therefore, a communication system offering ultra-high reliability while limiting the latency is needed. This thesis proposed a resource-efficient scheduling solution that would enable URLLC users with periodic and sporadic traffics to coexist, while satisfying the stringent requirements of URLLC. To represent the model in a realistic way, we assumed that users are classified into classes based on their packet generation probabilities.

So as to achieve the desired reliability, we proposed that a single packet should be transmitted multiple times; one copy through the highly reliable dedicated resources, and K copies over the shared pool. Most of the transmissions are carried out using the shared resources because the other alternative (i.e. having multiple transmission on dedicated resources) is uneconomical. The scheduling schemes considered for assigning dedicated resources are PS and GB, while the shared pool is accessed via the GF scheme. Both reliability and latency analyses were given for the three scheduling schemes. In order to utilize the resources efficiently, we assumed that users with periodic traffic access dedicated resources through PS, while sporadic-traffic users can access them through GB. This assumption was proven to reduce the amount of bandwidth required.

The main addition of this thesis to the available literature is the performance optimization of a network comprised of multiple classes of users distributed based on their packet generation rate. The objective is to find the optimal allocation of resources between the scheduling schemes that would result in either minimizing the required bandwidth or maximizing the traffic capacity, while satisfying the reliability requirements. Since users are divided into classes, a constant objective in all optimization problems is to find the optimal dividing class index that would make a separation line between PS-based and GB-based users. In order to get better insight on the effect of various parameters on the system, multiple optimization problems were implemented for different scenarios including finding the optimal packet length for a fixed number of information bits. Moreover, we showed the superiority of the proposed solution over the strategy of relying solely on the GF repetitions presented in the literature, in terms of the minimum amount of required bandwidth to meet the reliability target.

5.2 Future Work

We believe that the presented solution and results can be used as a reference for designing the scheduling protocol of a heterogenous system consisting of periodic and sporadic types of traffic. However, there is still room for improvement at the receiver side (i.e. base station in case of uplink). Future work will consider the impact of incorporating advanced receiver techniques on reliability. One of the effective ways to combat multipath fading and co-channel interference (collisions) is the use of Maximal Ratio Combining (MRC) which exploits antenna diversity at the receiver side. An MRC receiver equipped with multiple antennas combines the received signals as a weighted sum such that the output Signal to Interference plus Noise Ratio (SINR) is maximized [25]. This allows for the successful decoding of some of the collided packets in the shared pool. Even though this can increase the reliability of the system, it might result in extra latency due to the added complexity of the process of decoding a packet. Therefore, the additional processing delays at the receiver side should be considered.

References

[1] 3GPP TS 38.913 v 15.2.0, "Study on scenarios and requirements for next generation access technologies," Jul. 2018.

[2] M Series. IMT vision-framework and overall objectives of the future development of IMT for 2020 and beyond. *Recommendation ITU*, pages 2083–0, Sep. 2015.

[3] M. A. Siddiqi, H. Yu and J. Joung, "5G Ultra-Reliable Low-Latency Communication Implementation Challenges and Operational Issues with IoT Devices," *Electronics*, vol. 8, no. 9, p. 981, 2019.

[4] H. Zhang, N. Liu, X. Chu, K. Long, A. Aghvami, and V. C. M. Leung, "Network slicing based 5G and future mobile networks: Mobility, resource management, and challenges," *IEEE Commun. Mag.*, 55(8):138–145, Aug. 2017.

[5] B. Holfeld, D. Wieruch, T. Wirth, L. Thiele, S. A. Ashraf, J. Huschke, I. Aktas, and J. Ansari, "Wireless communication for factory automation: an opportunity for LTE and 5G systems," *IEEE Commun. Mag.*, 54(6):36–43, Jun. 2016.

[6] M. Luvisotto, Z. Pang and D. Dzung, "Ultra High Performance Wireless Control for Critical Applications: Challenges and Directions," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 3, pp. 1448-1459, June 2017.

[7] 5GPPP Association, "5G automotive vision," 5GPPP, White Paper, Oct. 2015.

[8] 5GPPP Association, "5G and e-health," 5GPPP, White Paper, Oct. 2015.

[9] T. Jacobsen, R. Abreu, G. Berardinelli, K. Pedersen, P. Mogensen, I. Z. Kovacs, and T. K. Madsen, "System Level Analysis of Uplink Grant-Free Transmission for URLLC," *IEEE Globecom Workshops (GC Wkshps)*, 2017.

[10]L. Buccheri, S. Mandelli, S. Saur, L. Reggiani, and M. Magarini, "Hybrid retransmission scheme for QoS-defined 5G ultra-reliable low-latency communications," *IEEE Wireless Communications and Networking Conference (WCNC)*, 2018.

[11]Z. Zhou, R. Ratasuk, N. Mangalvedhe, and A. Ghosh, "Resource Allocation for Uplink Grant-Free Ultra-Reliable and Low Latency Communications," *IEEE 87th Vehicular Technology Conference (VTC Spring)*, 2018.

[12]B. Singh, O. Tirkkonen, Z. Li, and M. A. Uusitalo, "Contention-Based Access for Ultra-Reliable Low Latency Uplink Transmissions," *IEEE Wireless Communications Letters*, vol. 7, no. 2, pp. 182–185, 2018.

[13]Y. Liu, Y. Deng, M. Elkashlan, A. Nallanathan and G. Karagiannidis, "Analyzing Grant-Free Access for URLLC Service," *IEEE Journal on Selected Areas in Communications*, pp. 1-1, 2020.

[14]Jacobsen, T., Abreu, R., Berardinelli, G., Pedersen, K., Mogensen, P., Kovacs, I. Z., & Madsen, T. K. (2017). "System Level Analysis of Uplink Grant-Free Transmission for URLLC," *IEEE Globecom Workshops (GC Wkshps)*, 2017.

[15]M. Deghel, P. Brown, S. E. Elayoubi, and A. Galindo-Serrano, "Uplink Contention-based Transmission Schemes for URLLC Services," *Proceedings of the 12th EAI International Conference on Performance Evaluation Methodologies and Tools*, 2019.

[16]R. Kotaba, C. N. Manchon, N. M. K. Pratas, T. Balercia, and P. Popovski, "Improving Spectral Efficiency in URLLC via NOMA-Based Retransmissions," *IEEE International Conference on Communications (ICC)*, 2019.

[17]S. Dogan, A. Tusha, and H. Arslan, "NOMA with Index Modulation for Uplink URLLC Through Grant-Free Access," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 6, pp. 1249–1257, 2019.

[18]M. Amjad and L. Musavian, "Performance Analysis of NOMA for Ultra-Reliable and Low-Latency Communications," *IEEE Globecom Workshops (GC Wkshps)*, 2018.

[19]3GPP TS 38.211 v15.7.0, "Physical channels and modulation," Sep. 2019.

[20]3GPP TS 38.101-1 v15.2.0, “User Equipment (UE) radio transmission and reception; Part 1: Range 1 Standalone,” Jul. 2018.

[21]C. E. Shannon, “A Mathematical Theory of Communication,” *Bell System Technical Journal*, vol. 27, no. 4, pp. 623–656, 1948.

[22]Y. Polyanskiy, H. V. Poor, and S. Verdu, “Channel Coding Rate in the Finite Blocklength Regime,” *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2307–2359, 2010.

[23]M. Chiani, D. Dardari and M. Simon, “New exponential bounds and approximations for the computation of error probability in fading channels,” *IEEE Transactions on Wireless Communications*, vol. 24, no. 5, pp. 840-845, 2003.

[24]3GPP TS 38.214 v 15.2.0, “Physical layer procedures for data,” Jul. 2018.

[25]G. Berardinelli, N. H. Mahmood, R. Abreu, T. Jacobsen, K. Pedersen, I. Z. Kovacs, P. Mogensen, “Reliability Analysis of Uplink Grant-Free Transmission Over Shared Resources”, *IEEE Access*, vol. 6, pp. 23602-23611, 2018.