Received 20 January 2023; revised 16 February 2023; accepted 2 March 2023. Date of publication 9 March 2023; date of current version 22 March 2023.

Digital Object Identifier 10.1109/OJCOMS.2023.3254816

Beyond 5G Resource Slicing With Mixed-Numerologies for Mission Critical URLLC and eMBB Coexistence

ALI ESMAEILY [1] (Student Member, IEEE), H. V. KALPANIE MENDIS [1], TOKTAM MAHMOODI [2] (Senior Member, IEEE), AND KATINA KRALEVSKA [1] (Member, IEEE)

[1] Department of Information Security and Communication Technology, Norwegian University of Science and Technology, 7491 Trondheim, Norway

[2] Department of Engineering, King's College London, WC2R 2LS London, U.K.

CORRESPONDING AUTHOR: A. ESMAEILY (e-mail: ali.esmaeily@ntnu.no)

ABSTRACT Network slicing has been a significant technological advance in the 5G mobile network allowing delivery of diverse and demanding requirements. The slicing grants the ability to create customized virtual networks from the underlying physical network, while each virtual network can serve a different purpose. One of the main challenges yet is the allocation of resources to different slices, both to best serve different services and to use the resources in the most optimal way. In this paper, we study the radio resource slicing problem for Ultra-Reliable Low Latency Communications (URLLC) and enhanced Mobile Broadband (eMBB) as two prominent use cases. The URLLC and eMBB traffic is multiplexed over multiple numerologies in 5G New Radio, depending on their distinct service requirements. Therein, we present our optimization algorithm, Mixed-numerology Mini-slot based Resource Allocation (MiMRA), to minimize the impact on eMBB data rate due to puncturing by different URLLC traffic classes. Our strategy controls such impact by introducing a puncturing rate threshold. Further, we propose a scheduling mechanism that maximizes the sum rate of all eMBB users while maintaining the minimum data rate requirement of each eMBB user. The results obtained by simulation confirm the applicability of our proposed resource allocation algorithm.

INDEX TERMS B5G, eMBB, numerology, puncturing, resource allocation, URLLC.

I. INTRODUCTION

THE NEXT-GENERATION mobile networks are intended to support the diverse requirements of the vertical industries, thus, to support a wide range of devices and applications. The fifth generation (5G) and beyond 5G (B5G) networks expand not only mobile broadband services compared to the fourth generation (4G) but also address new service-oriented use cases that involve innovative healthcare delivery, smart transportation systems, factory automation, and smart grids.

To address the diversity of applications and services served by 5G, the community categorized these services into three classes. These include massive Machine-Type Communications (mMTC), enhanced Mobile Broadband (eMBB), and Ultra-Reliable Low-Latency Communications (URLLC). mMTC is designed to provide connectivity for thousands of devices spread over a wide coverage. mMTC requires a low data rate and low power connectivity for enormous amounts of sensor/actuator devices (i.e., the Internet of Things (IoT)). eMBB deals with high data rates, high spectral efficiency, and low latency and can be considered as a direct extension of the 4G broadband services. URLLC necessitates significantly low End-to-End (E2E) latency and ultra-high reliability, and it is associated with tactile Internet [1]. URLLC is also referred to as mission-critical communications as it enables real-time control and automation of dynamic processes in various fields, such as industrial process automation and manufacturing, power distribution, or traffic management and safety.

The major challenge in providing such diverse and abundant services is that the physical infrastructure resources are scarce to meet all requirements. Thus, these resources need

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

to be deployed intelligently to deliver such services. In order to fulfill the above-specified diverse performance requirements imposed by B5G use cases, the next-generation mobile network has to be redesigned. Network slicing is a promising paradigm, which allows heterogeneous services to coexist within the same network architecture. The approach towards softwarization, virtualization, and cloudification as enabling technologies of network slicing has brought tremendous progress and benefits in terms of programmability, flexibility, and innovative ideas to service provisioning. Hence, network slicing leverages the benefits of a virtualized resource sharing environment enabled by Software-Defined Networking (SDN) and Network Function Virtualization (NFV) [2], [3], [4]. Based on softwarization and virtualization, it is capable of enabling Network-as-a-Service (NaaS) [5] and allows the coexistence of multiple networks on the same physical infrastructure. An E2E network slice is composed of the Radio Access Network (RAN), transport and Core Network (CN) sub-network slices in between the end (user) devices [6]. In this work, we consider slicing in the RAN, which is a constituting part of an E2E network slice.

5G New Radio (5G NR) follows the same principles of Orthogonal Frequency Division Multiple Access (OFDMA) technology which was adopted in Long Term Evolution (LTE) and LTE-Advanced (LTE-A). 5G NR supports multiple waveform configurations, which results in scalable numerologies. The resulting flexible frequency-time lattice is designed to support diverse requirements imposed by different traffic classes. URLLC users are typically mission-critical; therefore, they need to be prioritized over the eMBB users, which are typically considered best effort users. The coexistence of eMBB and URLLC users in the same mobile network is, hence, demanding given the trade-off between simultaneously achieving high data rates for the eMBB users and the ultra-reliability and low latency for the URLLC users.

A. RELATED WORK

Resource allocation and orchestration are vital aspects of network slicing as the logical E2E slices are realized upon a shared resource pool. To this end, numerous research works have considered radio resource allocation and proposed various scheduling algorithms. The 3rd Generation Partnership Project (3GPP) [7] has proposed a superposition/puncturing method for multiplexing URLLC and eMBB traffic in 5G cellular systems. The authors in [8] study the coexistence problem of eMBB and URLLC users in 5G networks. They formulate a joint resource allocation problem that can satisfy both eMBB user rate and URLLC interrupt probability requirements. They assign mini-slots for URLLC users and calculate the transmission power of URLLC users, ensuring the reliability constraint. A similar study is performed in [9], which also studies the resource slicing problem for 5G eMBB and URLLC services. The resource slicing problem is formulated as an optimization problem that aims at maximizing the eMBB data rate. The problem is subject to a URLLC reliability constraint while considering the variance of the eMBB data rate to reduce the impact of immediately scheduled URLLC traffic on the eMBB reliability. An optimization-aided deep reinforcement learning-based framework is proposed to solve the formulated problem.

The dynamic multiplexing scheme [7] is recognized as a promising technique to enable the coexistence of the eMBB and URLLC transmissions over the 5G NR and thus has attracted much attention in academia and industry. The authors in [10] evaluate the coexistence technique for eMBB and URLLC based upon a punctured scheme. They extend the study to formulate an optimization problem aiming to maximize the minimum expected achievable rate of eMBB User Equipment (UEs) while fulfilling the provisions of the URLLC traffic. In study [11], the radio resources are scheduled among the eMBB UEs on a time slot basis, whereas they are handled for URLLC UEs on a mini-slot basis. They use a penalty successive upper bound minimization-based algorithm for eMBB UEs, while the optimal transportation model is adopted to solve the same URLLC UEs problem. They also present a heuristic algorithm for efficient scheduling of PRBs among eMBB UEs.

Authors of [12] model the impact of the URLLC transmission over the scheduled eMBB traffic via loss functions caused by the URLLC traffic. The work in [13] analyzes the multiplexing of the eMBB and URLLC traffic in the Cloud-RAN (C-RAN) environment. The work in [14] investigates the performance trade-offs between eMBB and URLLC traffic types in a multi-cell C-RAN architecture under Non-Orthogonal Multiple Access (NOMA) and OMA access strategies. The work outcome reveals the advantage of employing the orthogonal-based solution for degrading the mutual interference of the eMBB and URLLC traffic. The authors also demonstrate the potential benefits of puncturing in improving the efficiency of fronthaul usage by discarding received mini-slots affected by URLLC interference. The authors in [15] present a puncturing scheme for transmitting low latency communication traffic, multiplexed on a downlink shared channel with eMBB. They also propose recovery mechanisms for the impacted eMBB users to minimize the capacity loss for eMBB users due to low latency communication traffic. A group of authors considers an optimal resource assignment under different channel conditions within a mixed numerology approach in [16], [17]. The work presented in [18] focuses on the scheduling problem for heterogeneous services within a mixed numerology approach aiming to maximize the number of satisfied users while meeting latency demand and data transmission requirements. Mini-slots enable transmissions that can be performed in a shorter time than the regular slot duration. In higher numerologies, the use of wider Sub-Carrier Spacings (SCSs) provides shorter slot durations. Consequently, low-latency communications can be enabled by combining mixed numerology and mini-slot approaches. 3GPP proposed mixed numerology with mini-slots that use single numerology with shorter slot durations than a regular slot for that predefined numerology

in order to support multiple services on the same carrier [19]. The work in [20] offers a model to optimize the numerology and resource allocation for mixed numerology systems, which employ the mini-slot approach.

The work in [21] aims to maximize the minimum expected achieved rate of eMBB users and fairness between them by employing a one-to-one matching game to compute appropriate eMBB and URLLC pairs for URLLC resource allocation. The authors of [22] and [23] aim at maximizing the aggregated throughput of the eMBB and URLLC users while mitigating the Inter-Numerology Interference (INI). They consider satisfying the minimum acceptable throughput of the eMBB and maximum allowed delay of the URLLC users according to their corresponding service requirements. The authors propose a deep reinforcement learning INI-aware agent to overcome the computation complexity of the optimization problem. Their method offers a spectrum allocation fulfilling the eMBB and URLLC service requirements while reducing the INI. Finally, they analyze their results delivered by the INI-aware agent when the URLLC traffic statistic is modeled based on mobile and industrial networks. Reference [24] formulates the RAN slicing problem between eMBB and URLLC users as a multi-timescale problem and proposes a hierarchical deep neural network algorithm to assign radio resources to their corresponding users. The authors model the selection of slice parameters within a time slot as a partially observable Markov decision process and present an algorithm to define configuration parameters for the eMBB and URLLC slices efficiently.

The work in [25], the authors compute the achievable latency for the industrial network scenario based on an accurate system-level simulation. Their primary focus is determining 5G NR configurations that are more relevant for Industry 4.0 applications to analyze the effect of reserving bandwidth for URLLC services. Reference [26] defines a context of the network based on combined statistical characteristics from the wireless channel and UEs' service requirements to train a Mondrian forest to predict an optimal mixed-numerology profile. The authors of [27] work on solving the challenges of radio resource allocation in the mmWave band of 5G NR by proposing a deep reinforcement learning-based scheduler. The scheduler allocates resources for a list of UEs to satisfy their different slice's SLA requirements according to the channel quality of each UE. Paper [28] presents a resource allocation strategy that combines latency, control channel, hybrid automatic repeat request, and radio channel quality in determining the transmission resources for different users. The approach minimizes the latency and bypasses unwarranted costly segmentation of URLLC payloads over several transmissions. Reference [29] addresses the problem of joint admission control and resource scheduling for URLLC by utilizing a standard continuous SNR model, where all allocated resource blocks contribute to the success probability, and a binary SNR model, where each resource block is classified as active or inactive according to a SNR threshold. In congestion cases, the work focuses on discovering a subset of users that can be scheduled at the same time.

The authors in [30] develop a joint optimization problem for power and bandwidth allocation with long-term conditions of queues backlog for the eMBB users. They utilize the Lyapunov drift-plus-penalty technique to create the relationship between the long-term constraints and the short-term optimization problem. Furthermore, they employ a one-to-one matching procedure to solve the slicing puncture problem. The work in [31] designs a coordinated multi-point multi-numerology network to improve the throughput of eMBB and latency of URLLC users. The authors solve a subcarrier and power allocation problem with the objective of maximizing the system sum rate. They show that their designed network has a higher sum data rate, lower delay, and throughput outage compared to the traditional non-coordinated multi-point single numerology scenarios. Reference [32] concentrates on minimizing the rate loss of the eMBB users and packet segmentation loss of URLLC users while fulfilling the QoS requirements of eMBB and URLLC use cases. They consider the case of one-to-one pairing in which one URLLC packet can be paired with only one eMBB. They employ a bi-level optimization problem that includes one inner and one outer problem. The inner problem seeks to discover the optimal power and frequency resources for each URLLC and eMBB pair, and the outer problem desires to search for the optimal eMBB-URLLC pairing policy. They also generalize the problem for many-to-many pairing while undervaluing the overhead due to URLLC packet segmentation.

The authors in [33] aim at minimizing the decoding error rate of URLLC users while ensuring the demand for the throughput of eMBB users. They propose a block coordinate descent optimization algorithm to obtain the optimal bandwidth allocation, puncture weight, and transmit power. Paper [34] focuses on studying eMBB and URLLC use cases in networks that are assisted by a Reconfigurable Intelligent Surface (RIS). The authors jointly optimize the power and frequency allocation problem and the RIS phase shift matrix to enhance the eMBB sum rate and URLLC reliability. The work in [35] concentrates on eMBB and URLLC use cases in a massive MIMO system by providing a unified information-theoretic framework incorporating an infinite-blocklength analysis of the eMBB spectral efficiency with a finite-blocklength analysis of the URLLC error probability. The work relies on the use of mismatched decoding and saddlepoint approximation.

Compared to the works presented above, in our previous work [36], we maximize the data rate for each of the eMBB users while guaranteeing a minimum acceptable data rate requirement per eMBB user. We develop the resource allocation problem by formulating a loss function for each eMBB user that experiences an adverse impact on its data rate due to the puncturing by the incoming URLLC traffic. We aim to minimize such negative impact of URLLC traffic upon eMBB users by introducing a puncturing rate threshold. In

TABLE 1. Comparison of related work and the proposed work in this paper for eMBB and URLLC coexistence, where ✓ denotes that the corresponding work covers the topic and ✗ denotes that the corresponding work does not cover the topic.

Related Work	Puncturing method	Mixed Numerology	URLLC latency	URLLC reliability	URLLC traffic classification	eMBB Loss function definition	Power allocation	eMBB data rate	Resource Block allocation	Channel state
[8]	✓	✗	✗	✓	✗	✗	✓	✓	✓	✗
[9]	✓	✗	✗	✓	✗	✗	✓	✓	✓	✓
[10]	✓	✗	✓	✓	✗	✓	✗	✓	✓	✗
[11]	✓	✗	✓	✓	✗	✓	✗	✓	✓	✗
[12]	✓	✗	✗	✗	✗	✓	✗	✓	✓	✓
[13]	✗	✗	✓	✗	✗	✗	✓	✓	✗	✓
[14]	✓	✗	✓	✗	✗	✗	✓	✓	✗	✓
[15]	✓	✗	✓	✗	✗	✗	✗	✓	✗	✓
[16]	✗	✓	✗	✗	✗	✗	✗	✓	✗	✓
[17]	✗	✓	✗	✗	✗	✗	✓	✓	✗	✓
[18]	✗	✓	✗	✗	✗	✗	✗	✓	✓	✗
[20]	✗	✓	✓	✗	✗	✗	✗	✓	✗	✓
[21]	✓	✓	✓	✓	✗	✗	✗	✓	✓	✓
[22], [23]	✗	✓	✓	✓	✗	✗	✗	✓	✓	✓
[24]	✓	✓	✓	✓	✗	✗	✓	✓	✓	✓
[25]	✗	✓	✓	✗	✗	✗	✗	✓	✗	✗
[26]	✗	✓	✓	✗	✗	✗	✗	✓	✗	✗
[27]	✗	✓	✓	✗	✗	✗	✗	✓	✓	✗
[28]	✗	✗	✓	✓	✗	✗	✗	✓	✓	✗
[29]	✗	✗	✓	✓	✗	✗	✗	✗	✓	✗
[30]	✓	✗	✓	✗	✗	✓	✓	✓	✓	✓
[31]	✗	✓	✓	✗	✗	✗	✓	✓	✓	✓
[32]	✓	✗	✓	✓	✗	✓	✓	✓	✗	✓
[33]	✓	✗	✗	✓	✗	✓	✓	✓	✓	✗
[34]	✓	✗	✓	✓	✗	✓	✓	✓	✓	✓
[35]	✓	✗	✗	✗	✗	✗	✓	✓	✗	✓
[36]	✓	✗	✓	✓	✗	✓	✓	✓	✓	✓
Our work	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

this paper, we extend our previous work by incorporating mixed numerologies for different URLLC traffic classes. We also introduce the MiMRA algorithm for the resource-slicing problem between URLLC and eMBB traffic. Some of the recent related works, such as [22], [23], [30], [32], and [34], address the main technical challenges in the eMBB and URLLC coexistence problem. Nevertheless, no work considers URLLC traffic classification. Although all of the URLLC use cases share low latency and ultra-high reliability characteristics, each specific use case holds its own distinct and exclusive value, such as the Motion control use case compared to the Closed-loop control use case as discussed further in Section II-B. Such URLLC use cases also need prioritization in serving by the network. Thus, classifying URLLC traffic is crucial. Besides, each specific URLLC class holds a different packet size, and this feature becomes extra-critical when transmitting such packets promptly. Moreover, considering accurate power allocation to the eMBB and URLLC users is also vital in order to fulfill their service requirements while overcoming noise and interference. Consequently, there is a need for simultaneously addressing puncturing over mixed numerologies while differentiating the URLLC traffic. This motivates our contributions in this paper, outlined in the next section. Table 1

summarizes the comparison of the related work and our proposed work.

B. CONTRIBUTIONS

The main contributions of this paper are as follows:

- We describe the radio resource allocation problem for the coexistence of eMBB and URLLC traffic scheduling by employing the puncturing method over different numerologies. We formulate the resource allocation problem according to the eMBB resource block allocation, eMBB power allocation, and scheduling of different URLLC traffic classes. Our main objective is to maximize the *sum rate* of the eMBB users while fulfilling the minimum acceptable data rate of each eMBB user in order to deliver fairness in allocating radio resources. Concurrently, the resource allocation problem has to satisfy the extra low latency and ultra-high reliability requirements of the URLLC users.
- We categorize the URLLC traffic into different classes. Each class represents a portion of the traffic that has been generated by the URLLC users belonging to a particular URLLC use case. To the best of our knowledge and following Table 1, there is no similar work that investigates together the puncturing method and mini-slots with 5G NR mixed-numerology to fulfill distinct URLLC classes' requirements (extra low latency and ultra-high reliability) on the one hand and to maximize the sum rate of the eMBB users on the other hand. In this way, apart from eMBB users, we can also differentiate and prioritize URLLC traffic classes as they belong to various URLLC use cases and thus hold different QoS requirements.
- We define precisely a loss function of the eMBB user's data rate to capture the impact of puncturing by the overlapped traffic of each URLLC class according to the number and size of the URLLC packets within each class. The loss function is expressed per TTI in each specific numerology and for every particular radio resource allocated to each eMBB user.
- We propose an optimization strategy called Mixed-numerology Mini-slot based Resource Allocation (MiMRA) that guarantees the loss in eMBB data rate due to the overlapped URLLC traffic is minimal. Consequently, the achievable data rate for the eMBB users is not impacted immensely. Furthermore, we represent a *puncturing rate threshold* to limit the such impact.

C. ORGANIZATION

The remainder of this paper is organized as follows. In Section II we present a few fundamental concepts that are related to this work. Section III explains the system model of our network. In Section IV, we describe the proposed optimization strategy for eMBB/URLLC coexistence. In Section V, we illustrate the numerical results of the analysis. Finally, Section VI concludes the paper.

II. PRELIMINARIES

In this section, we discuss a few fundamental elements upon which the work of this paper is built.

A. 5G NEW RADIO

5G New Radio (NR) is designed to support deployment across a wide range of frequencies. Two different frequency ranges are designated for 5G NR named: Frequency Range 1 (FR-1) and Frequency Range 2 (FR-2) [37]. The bands in FR-1 are envisaged to carry much of the traditional cellular mobile communications traffic. The higher frequency bands in the range FR-2 aim to provide short range very high data rate capability for the 5G radio. Thus, 5G NR can operate in both the sub-6 GHz bands, some of which are traditionally used by previous standards, as well as millimeter wave (mmWave) bands with a shorter range but higher available channel bandwidths.

1) 5G SCALABLE NUMEROLOGIES

Distinct from LTE-A, 5G NR supports multiple waveform configurations referred to as numerologies. A numerology represents a set of parameters such as SCS, OFDM symbol length, and Cyclic Prefix (CP). LTE supports carrier bandwidths of up to 20 MHz with a mainly fixed OFDM numerology (15 KHz SCS). Nevertheless, NR is designed to offer scalable OFDM numerologies to support diverse spectrum bands and deployment models. This is achieved by creating multiple numerologies formed by scaling the basic LTE SCS with 2^{μ}, where μ is an integer between 0 and 4. The numerology is selected independently of the frequency band, with possible SCS of 15 KHz to 240 KHz. Regardless of the numerology, the length of a radio frame and a subframe are always 10 ms and 1 ms, respectively. The difference is the number of time slots within a subframe and the number of symbols within a time slot.

Table 2 presents the main features of each of the five numerologies defined in 5G NR [38]. The following is the terminology used in this paper.

- *Numerology:* A numerology represents a set of parameters such as SCS, OFDM symbol length, and CP.
- *Frame:* Similar to LTE, 10 subframes, each lasting for $1ms$ construct one frame.
- *Slot:* A slot consists of 14 OFDM symbols and is transmitted within a transmission time interval (TTI).
- *Transmission time interval (TTI)/(eMBB) time slot:* Corresponds to 1 subframe duration ($1ms$) that is required to encapsulate non delay-sensitive data (transport blocks) from higher radio protocol stack layers and deliver it to the physical layer in order to transmit it via the radio interface.
- *Resource Block (RB):* In this paper, a RB in 5G NR is defined as 12 consecutive subcarriers in the frequency domain and 14 symbols in the time domain. With different sizes of slots and subcarriers of different numerologies, the size of the RB may change, as illustrated in Figure 1.

TABLE 2. 5G new radio numerologies [38].

Numerology, μ	SCS [KHz]	#symbols per slot	#slots per subframe	Cyclic prefix (CP)	Symbol duration [μs]	CP duration [μs]
0	15	14	1	Normal	71.43	4.69
1	30	14	2	Normal	35.71	2.34
2	60	14, 12	4	Normal, extended	17.86	1.17
3	120	14	8	Normal	8.92	0.57
4	240	14	16	Normal	4.46	0.29

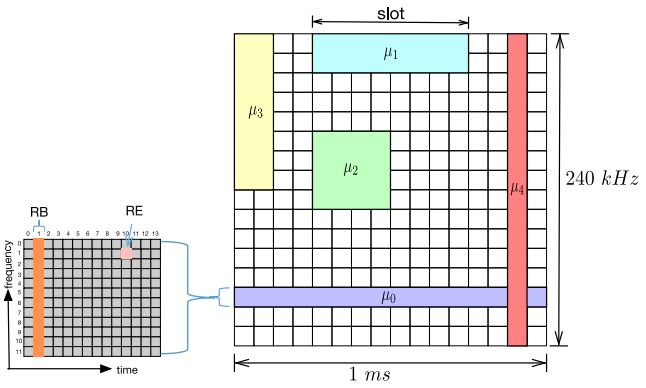


FIGURE 1. 5G flexible numerology structures.

- *Resource Element (RE):* RE is the smallest unit within the 5G NR resource grid, consisting of one subcarrier in the frequency domain and one OFDM symbol in the time domain.
- *Cyclic Prefix (CP):* CP is required to eliminate Inter-Symbol Interference (ISI) due to multipath signals. 5G NR supports both normal CP and extended CP. With a normal CP, each slot is formed by 14 OFDM symbols, while 12 are only available when using an extended CP.

5G NR can support a wide range of services, devices, and deployments. Another new concept in 5G NR is called Bandwidth Part (BWP). A BWP is a set of contiguous RBs configured inside a channel bandwidth; thus, the width of a BWP may be smaller than or equal to the channel bandwidth. The motivation behind introducing BWP is such that it could be challenging to use the larger 5G bandwidths for the UEs and to support UE bandwidth adaptation to help reduce device power consumption [39]. Besides, BWPs can be of various numerologies, which can be employed to decrease the latency for particular services, such as in URLLC use cases.

Employing multiple numerologies in the 5G NR enhances the flexibility of scheduling use cases with diverse service requirements via performing slicing in the RAN. With a shorter duration of slots, transmissions can be scheduled much faster than a traditional LTE-based network. Furthermore, NR enables both uplink and downlink transmissions within a slot, making it possible to support low latency traffic. In addition, different numerologies support multiple deployment scenarios from sub-1 GHz range to mmWave applications. The higher numerologies $\mu = 3$ and $\mu = 4$ support high frequencies in the mmWave range defined in the range of FR-2. Furthermore, since the symbol length and SCS are inversely proportional to each other, wider SCSs reduce the CP length, which is an overhead to a system. This is especially useful for smaller cells where delay spread is low. For applications that tolerate longer delay spread, narrower SCSs are preferable. In the work [40], the authors present a criterion for selecting the most preferable numerology for a set of services under given network conditions.

However, the flexibility provided by the multi-numerology structure of 5G NR introduces a non-orthogonality into the system. This causes interference between the multiplexed numerologies, known as Inter-Numerology Interference (INI). Several papers analyze INI and present various INI management techniques [22], [23], [41], [42], [43].

2) 5G MINI-SLOT

A slot is a unit for transmission used by the gNB scheduling mechanism. Typically, a slot occupies either 14 (normal CP) or 12 (extended CP) OFDM symbols (see Table 2). 5G NR allows transmissions with a flexible start position and a duration shorter than a regular slot duration, which is referred to as a mini-slot. A mini-slot is the minimum scheduling unit used in 5G NR. Though, in principle, a mini-slot can be as short as one OFDM symbol in Release 15, mini-slots are limited to 2, 4, and 7 OFDM symbols [44]. In low-latency scenarios such as URLLC, a transmission needs to begin immediately without waiting for the start of a slot. Mini-slot transmission facilitates achieving lower latency in 5G NR architecture and minimizes interference to other RF links. Unlike slots, mini-slots are not tied to the frame structure. This helps in puncturing the existing frame without waiting to be scheduled.

B. URLLC AND EMBB USE CASES

A vertical domain is an industry or group of enterprises in which similar products or services are developed, produced, and provided. The operation of a vertical industry is hugely interdependent on an Information and Communications Technology (ICT) infrastructure. Depending on the products or the services they deliver, the underlying ICT infrastructure must be designed in a way that it can accommodate all

TABLE 3. Communication service performance requirements for different URLLC use cases [45].

Use case	availability (%)	reliability (MTBF)	maximum E2E latency	data size (Bytes)
Motion control in factory automation	$99,9999$	10 years	$1ms - 500ns$	-
Closed-loop control in process automation	$99,9999$	> 1 year	$< 10ms$	-
Fault location, isolation, and restoration (FLISR)	$99,9999$	-	$< 5ms$	< 1500
Wind power plant network	$99,9999999$	10 years	$16ms$	-

service requirements demanded by such vertical industries. The vertical domains addressed in this paper are:

- *Power distribution:* Modern society is highly dependent on the reliability and resiliency of the power grid. The energy sector is currently subject to a fundamental change caused by the evolution toward renewable energy, i.e., an increasing number of power plants based on solar and wind power. These changes lead to bi-directional electricity flows and increased dynamics of the power system. New sensors and actuators are being deployed in the power system to efficiently monitor and control the volatile conditions of the grid, requiring real-time information exchange. The emerging power distribution grid is also referred to as the smart distribution grid. The smartness enhances insight into both the grid as a power network and the grid as a system of systems.
- *Power generation:* This domain comprises all aspects of centralized power generation, i.e., the centralized conversion of chemical energy and other forms of energy into electrical energy. Examples of pertinent systems are large gas turbines, steam turbines, combined-cycle power plants, and wind farms. In addition, the planning and installation of respective equipment and plants, as well as the operation, monitoring, and maintenance of these plants are encompassed by this vertical domain.
- *Manufacturing:* The manufacturing industry is currently subject to a fundamental change due to the fourth industrial revolution (Industry 4.0). It requires improvements in flexibility, versatility, resource efficiency, cost efficiency, worker support, and quality of industrial production and logistics in order to address the needs of increasingly volatile and globalized markets.

In the following, we present several use cases within selected 5G vertical industries requiring URLLC or eMBB communication services, which some of them will be considered later in a scenario for our simulations.

1) URLLC USE CASES

- *Factory automation in manufacturing:* Factory automation serves the automated control, monitoring, and optimization of processes in a factory. It deals with applications such as closed-loop control, motion controllers, robotics, and computer-integrated manufacturing. Factory automation is a key enabler for industrial mass production with high quality and cost-efficiency. Thus, related applications are characterized by strict requirements on the underlying communication infrastructure regarding availability and latency.
- *Process automation in manufacturing:* In the closed-loop control use case for process automation, several sensors are installed in a plant, and each sensor performs continuous measurements. The measurement data is transported to a controller, which takes a decision to set actuators. The latency and determinism in this use case are crucial. Therefore, this use case has very stringent requirements in terms of latency and service availability.
- *Fault Location, Isolation and Service Restoration (FLISR) in power distribution:* The FLISR is an essential operation to support the self-healing of power distribution grids. Typically, in power distribution grids, each feeder section has a controller device. Using Peer-to-Peer (P2P) communication among the Intelligent Electronic Devices (IEDs), the system operates autonomously without the intervention of the control center. In P2P communication via IEC 61850 GOOSE (Generic Object-Oriented Substation Event) messages are sent periodically (in steady-state) by each IED to neighboring IEDs of the same feeder and are not acknowledged. The data rate per IED is low in steady-state, but GOOSE bursts with high data rates occur, especially during fault situations, and require low E2E latency and high reliability.

Table 3 presents the different communication service performance requirements for different URLLC use cases mentioned above.

2) EMBB USE CASES

- *Remote grid surveillance in power distribution:* Critical infrastructures such as power distribution grids must be continuously monitored and controlled. Such critical infrastructures are heavily exposed to threats posed by malicious actors as well as potentially catastrophic natural disasters. As a result, there is a trend for smart distribution grids to incorporate video, photography, Unmanned Aerial Vehicles, and drones for visual surveillance for the supervision and observation of grid equipment.
- *Augmented (AR) or Virtual Reality (VR):* Use cases and applications also exist that require very high data rates

as offered by eMBB, such as augmented or virtual reality. Cloud-based AR/VR is the key technology enabling games, education, training.

- *Video streaming from event venues:* One potential application is large spectator events such as sports games or concerts, where spectators are located far away from the physical event location but are able to experience it, for instance, via live video streaming through social media. Also, the spectators are able to experience a front-row view of the action despite their physical location as a benefit of VR.

C. URLLC AND EMBB COEXISTENCE STRATEGIES

The incoming URLLC packets to a gNB have to be immediately sent through the scheduled eMBB transmissions and cannot be queued due to the strict latency requirements of URLLC traffic. The conventional orthogonal-based radio resource allocation mechanism is not suited for the coexistence of URLLC and eMBB traffic [46]. 3GPP defines two approaches for the coexistence of these heterogeneous services with distinct requirements.

1) DYNAMIC MULTIPLEXING

The superposition or puncturing scheme is one of the proposals from 3GPP to efficiently multiplex eMBB and URLLC data transmissions via the 5G NR [7]. eMBB traffic is scheduled at the beginning of slots. URLLC packets may arrive during an ongoing eMBB transmission, and URLLC traffic can be immediately overlapped at any mini-slot. If eMBB transmissions are allocated zero power when URLLC traffic is overlapped, then it is referred to as the puncturing of eMBB transmissions. If the gNB chooses non-zero transmission powers for both eMBB and overlapping URLLC traffic, that is referred to as the superposition of URLLC traffic over eMBB traffic. It is worth mentioning that there is a tradeoff between employing superposition instead of puncturing. Utilizing superposition will enhance the performance in terms of the eMBB sum rate. Nevertheless, this advantage comes with the cost of 1) eMBB user's interference over URLLC user resulting in increasing the risk of violating the URLLC reliability requirement, and 2) computational complexity in URLLC users due to performing the Successive Interference Cancellation (SIC) technique [47]. Besides, there is no guarantee of delivering fairness in allocating resources between eMBB users since the objective of superposition is to improve the eMBB sum rate and not necessarily to fulfill the minimum acceptable data rate of each eMBB user. There has been a solution by allocating more power to the URLLC user, compared to the eMBB user, in order to reduce bit-error-rate and therefore higher reliability, and eliminate using SIC in the URLLC user [48]. However, in this solution, it is assumed that the gNB allocates more power to the URLLC user. This method is against one of our objectives, as operating such a method results in disregarding accurate and optimum power allocation between eMBB and URLLC users. Another solution is employing the superposition or puncturing technique according to the gNB decision. This approach may not be feasible either, as URLLC traffic needs to be transmitted immediately. Due to decision time, switching time, and processing time between conducting superposition or puncturing technique by the gNB, employing such an approach can violate the low latency requirement of URLLC packets.

2) ORTHOGONAL SCHEDULING

The gNB pre-reserves a number of frequency channels for URLLC traffic. Two reservation mechanisms fall under the orthogonal scheduling; semi-static reservation and dynamic reservation [9]. In the semi-static scheme, the gNB intermittently broadcasts the frame structure configurations. However, in the dynamic reservation, the frame structure information is updated frequently and dynamically using the control channel of a scheduled user. The downside of this approach is that resources reserved for URLLC will be wasted in case there is no URLLC transmission. Furthermore, the dynamic scheme needs additional control overhead compared to the semi-static scheme.

D. ASSUMPTIONS

Several assumptions are considered in the problem formulation.

- We assume eMBB and URLLC downlink transmissions with different service requirements in terms of data rate, latency, and packet size. The URLLC traffic is coming from several URLLC priority classes. Each URLLC priority class contains data flow of a certain number of URLLC UEs that generate packets with a specific incoming rate (high and medium compared to the eMBB users), and they have a particular delay requirement.
- We focus on the dynamic puncturing of allocated resources to eMBB users by overlapping the URLLC traffic on the same radio resources.

III. SYSTEM MODEL

In this section, we explain the system model, we formulate the problem and we also present the proposed algorithm for eMBB/URLLC coexistence. Table 4 summarizes the notation used in this paper.

A. SYSTEM MODEL AND PROBLEM FORMULATION

We analyze and study downlink eMBB and URLLC traffic, i.e., transmitting traffic from a single gNB, that can operate with single or multiple antennas $j \in \mathcal{J} = \{1, 2, \ldots, J\}$, to User Equipment (UEs). For the sake of simplicity, we consider single antenna eMBB and URLLC UEs to envision Massive MIMO scenarios, as assumed in [49].

The gNB schedules the eMBB and URLLC traffic and transmits the corresponding data for each service type via its antennas towards eMBB and URLLC users over flat independent and identically distributed (i.i.d.) Rayleigh

TABLE 4. List of parameters used in the paper.

Notation	Meaning
$\mathcal{I}$	Set of URLLC traffic classes
$\mathbb{\mu}$	Set of numerologies
$\mathcal{J}$	Set of antennas
$\mathcal{K}$	Set of eMBB UEs
$\mathcal{N}$	Set of URLLC UEs
L	Maximum number of iterations
B_μ	Set of RBs in numerology μ
$\Delta f_{(\mu=\chi)}$	SCS of numerology $\mu=\chi$
m_μ	Mini-slot within the numerology μ
b_μ	RB b within the numerology μ
t_μ	Time slot t of numerology μ
$R^{eMBB}_{k_\mu}(t_\mu)$	Achievable rate in numerology μ for eMBB user k_μ at the time slot t_μ
$\phi^{eMBB}_{k_\mu}(t_\mu)$	Total amount of radio resources allocated to the eMBB user k_μ at time slot t_μ
$\gamma^{eMBB}_{k_\mu}(t_\mu)$	Total loss function (fraction of punctured RBs b_μ allocated to eMBB user k_μ at time slot t_μ)
$x_{k_\mu b_\mu}(t_\mu)$	Resource allocation coefficient for eMBB user k_μ
f_{b_μ}	Bandwidth of the RB b in numerology μ
$p^j_{k_\mu b_\mu}(t_\mu)$	Transmission power from the antenna j of the gNB over the RB b_μ to the eMBB user k_μ at time slot t_μ
$p^j_{n_{i\mu} b_\mu}(t_\mu)$	Transmission power from the antenna j of the gNB over the RB b_μ to the URLLC user n_i in time slot t_μ
$h^j_{k_\mu b_\mu}(t_\mu)$	Rayleigh fading channel gain of the transmission from the antenna j of the gNB over the RB b_μ to the eMBB user k_μ at time slot t_μ
$h^j_{n_{i\mu} b_\mu}(t_\mu)$	Rayleigh fading channel gain of the transmission from the antenna j of the gNB over the RB b_μ to the URLLC user n_i in time slot t_μ
$\sigma^2_{Total_{k_\mu}}$	Total interference and noise power impacts eMBB user k_μ
$\sigma^2_{Total_{n_{i\mu}}}$	Total interference and noise power impacts URLLC user $n_{i\mu}$
$D^i_{m_\mu,n_i}(t_\mu)$	Random variable indicating the number of incoming URLLC packets generated by n_i user in mini-slot m_μ
$\eta^i_{m_\mu,n_i}(t_\mu)$	Instantaneous packet size of URLLC UE $n_i \in \mathcal{N}_i = \{1,2,...,N_i\}$ belonging to the class i in the mini-slot m_μ of time slot t_μ
$D_{total}(t_\mu)$	Total incoming URLLC traffic at time slot t_μ
π_{k_μ}	Number of punctured mini-slots of eMBB user k in numerology μ
$th^{eMBB}(t_\mu)$	Puncturing rate threshold
θ^i_{max}	Outage probability threshold of the URLLC class i
R_{min}	Minimum acceptable eMBB data rate

fading channels. The gNB serves $k \in \mathcal{K} = \{1,2,\ldots,K\}$ total number of eMBB and $n \in \mathcal{N} = \{1,2,\ldots,N\}$ total number of URLLC UEs within a set of numerologies $\mu \in \mathbb{\mu} = \{0,\ldots,4\}$. Figure 2 illustrates a symbolic puncturing mechanism for the coexistence of eMBB and URLLC traffic classes for the numerologies $\mu=0$, $\mu=1$, and $\mu=2$ with $\Delta f_{(\mu=0)} = 15$ KHz, $\Delta f_{(\mu=1)} = 30$ KHz, and $\Delta f_{(\mu=2)} = 60$ KHz, $f_{b_{\mu=0}} = 180$ KHz, $f_{b_{\mu=1}} = 360$ KHz, and $f_{b_{\mu=2}} = 720$ KHz, $\text{TTI}_{(\mu=0)} = 1$ ms, $\text{TTI}_{(\mu=1)} = 0.5$ ms, and $\text{TTI}_{(\mu=2)} = 0.25$ ms, respectively. Within each specific numerology μ,

- The time domain is split into equally spaced time slots (TTIs) for the eMBB UEs' transmissions. The time slot is then subdivided into a fixed number of M_μ equally spaced mini-slots (short TTIs) where $m_\mu \in \mathcal{M}_\mu = \{1,2,\ldots,M_\mu\}$ denotes a mini-slot within the numerology μ.
- The radio resources in the frequency domain are divided into $b_\mu \in \mathcal{B}_\mu = \{1,2,\ldots,B_\mu\}$ RBs. Each RB b_μ contains 12 sub-carriers in the frequency domain and 14 OFDM symbols in the time domain.
- We refer to each eMBB user as k_μ, since depending on the gNB configuration, each eMBB user can be served via different numerologies in the various corresponding time slots.

According to the incoming arrival rates, the latency, and reliability requirements of different URLLC use cases mentioned in Table 3, the URLLC UEs are sub-categorized into different traffic classes $i \in \mathcal{I} = \{1,2,\ldots,I\}$. In each class i, a subset number of URLLC UEs N_i generate a traffic volume within mini-slot m_μ. Since there is no strict latency requirement for serving the eMBB users, the RBs are allocated to them at the beginning of each time slot. However, the sporadic URLLC requests can arrive at any time within a time slot, and due to the extreme latency requirement of such requests, the gNB needs to serve them immediately in a mini-slot instead of waiting for the next time slot. Therefore, the gNB punctures previously scheduled eMBB transmissions in mini-slots by applying zero power to these transmissions to serve the URLLC requests promptly.

The sporadic URLLC traffic impacts the previously scheduled eMBB users with the allocated radio resources in some mini-slots. Suppose there are two different URLLC traffic classes, i and $i+1$. Let us assume that these URLLC traffic classes arrive at the first mini-slot of the first time slot for the three numerologies $\mu=0$, $\mu=1$, and $\mu=2$, as it is shown in Figure 2. gNB determines to map the URLLC traffic to the eMBB UEs 1, 3, and 5, i.e., $k=1$, $k=3$, and $k=5$. In particular, the URLLC traffic of class i and $i+1$ punctures 1) the $m_0=1$ of the first slot of $k=1$ with 2 OFDM symbols per mini-slot, 2) the $m_1=1$ of the first slot of $k=3$ with 7 OFDM symbols per mini-slot, and 3) the $m_2=1$ of the first slot of $k=5$ with 7 OFDM symbols per mini-slot. The same idea is repeated for the last mini-slots of the eMBB UEs $k=2$, $k=4$, and $k=8$ in which URLLC traffic classes i and $i+1$ arrive randomly, and gNB determines to puncture them. The URLLC traffic of class i and $i+1$ punctures 1) the $m_0=7$ of the first slot of $k=2$ with 2 OFDM symbols per mini-slot, 2) the $m_1=2$ of the first slot of $k=4$ with 7 OFDM symbols per mini-slot, and 3) the $m_2=2$ of the first slot of $k=8$ with 7 OFDM symbols per mini-slot. The idea is that the generated URLLC packets belonging to different URLLC

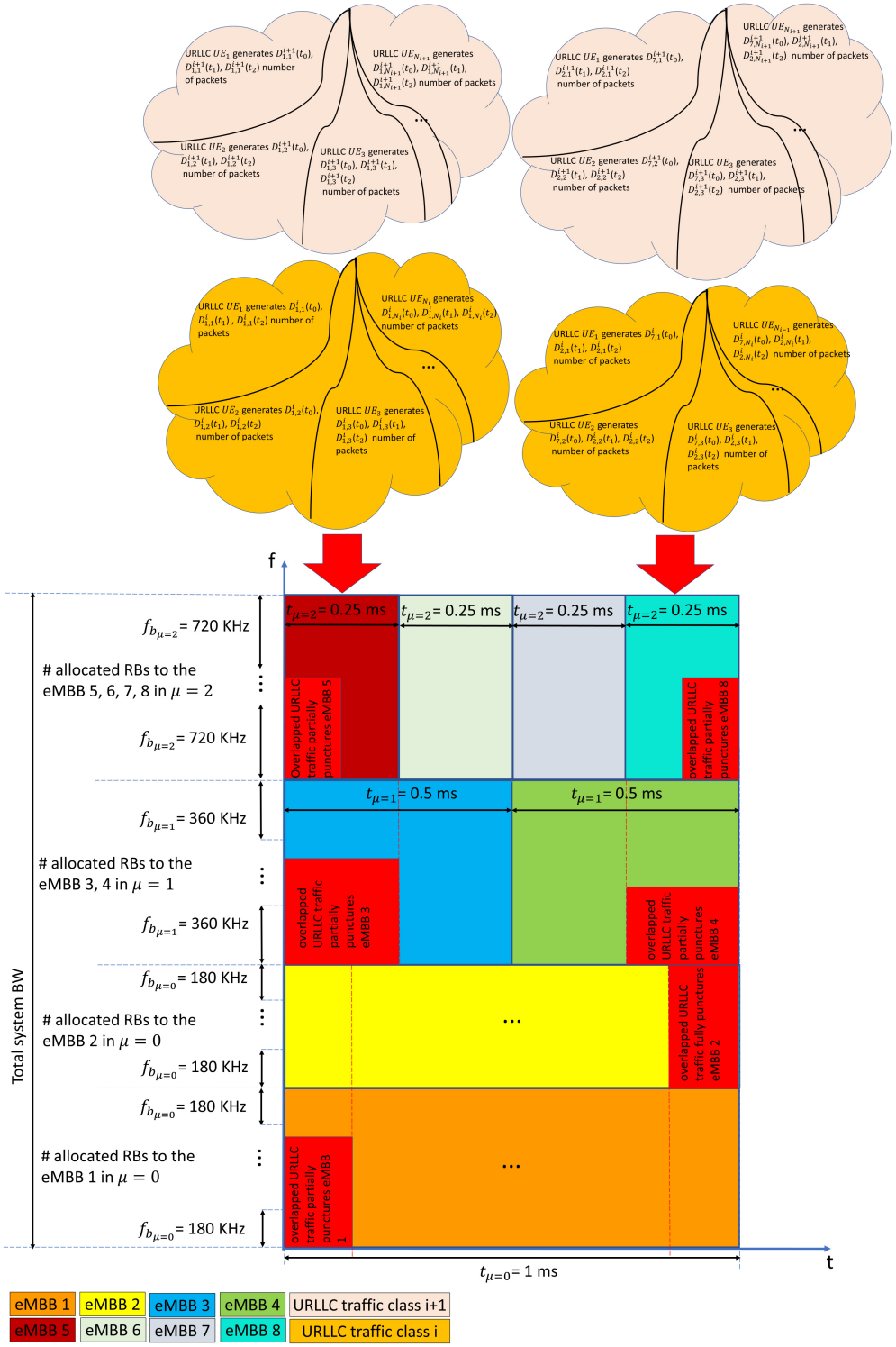


FIGURE 2. Coexistence of eMBB and URLLC traffic classes in downlink via the puncturing mechanism for $\mu = 0$, $\mu = 1$, and $\mu = 2$ numerologies.

classes are served. Hence, some of the allocated resources to the $k = 1, 2, 3, 4, 5, 8$ are punctured by the overlapped URLLC traffic.

Accordingly, the maximum achievable rate in the particular numerology μ for an eMBB user k_μ at the time slot t_μ over the whole allocated RBs can be formulated as follows:

$$R_{k_\mu}^{eMBB}(t_\mu) = \left[\phi_{k_\mu}^{eMBB}(t_\mu) - \gamma_{k_\mu}^{eMBB}(t_\mu)\right] \times R_{k_\mu,peak}^{eMBB}(t_\mu) \quad (1)$$

where the $\phi_{k_\mu}^{eMBB}(t_\mu)$ is the total amount of radio resources allocated to the eMBB user k_μ at time slot t_μ, $\gamma_{k_\mu}^{eMBB}(t_\mu)$ is the total loss function which indicates the fraction of punctured resources allocated to eMBB user k_μ at time slot

t_μ due to the incoming URLLC requests, and $R^{eMBB}_{k_\mu,peak}(t_\mu)$ is the peak achievable data rate of the eMBB user k_μ at time slot t_μ. This formulation is general, and by following the Shannon channel capacity, it can be further extended to:

$$R^{eMBB}_{k_\mu}(t_\mu) = \sum_{b_\mu=1}^{B_\mu}\left[\left(x_{k_\mu b_\mu}(t_\mu) f_{b_\mu} - \gamma^{eMBB}_{k_\mu b_\mu}(t_\mu)\right) \times log_2\left(1 + \frac{\sum_{j=1}^{J} p^j_{k_\mu b_\mu}(t_\mu) h^j_{k_\mu b_\mu}(t_\mu)}{\sigma^2_{Total_{k_\mu}}}\right)\right] \tag{2}$$

where $x_{k_\mu b_\mu}(t_\mu)$ is the resource allocation coefficient, $x_{k_\mu b_\mu}(t_\mu) = 1$ denotes that the RB b_μ is allocated to the eMBB user k_μ at time slot t_μ and $x_{k_\mu b_\mu}(t_\mu) = 0$ shows no allocation; f_{b_μ} is the bandwidth of the RB b_μ; $p^j_{k_\mu b_\mu}(t_\mu)$ is the transmission power from the antenna j of the gNB over the RB b_μ to the eMBB user k_μ at time slot t_μ; $h^j_{k_\mu b_\mu}(t_\mu)$ is the Rayleigh fading channel gain of the transmission from the antenna j of the gNB over the RB b_μ to the eMBB user k_μ at time slot t_μ; $\sigma^2_{Total_{k_\mu}} = \sigma^2_{ICI_\mu} + \sigma^2_{ISI_\mu} + \sigma^2_{Cher_{k_\mu}} + \sigma^2_{INI_{k_\mu}} + \sigma^2_{Noise}$ indicates the summation of Inter-carrier Interference (ICI), Inter-symbol Interference (ISI), Channel estimation error (Cher), INI, and noise power, respectively [20], [50]; and finally, $\gamma^{eMBB}_{k_\mu b_\mu}(t_\mu)$ indicates the fraction of punctured RB b_μ that is allocated to eMBB user k_μ at time slot t_μ.

Now, in each specific numerology μ, we consider $D^i_{m_\mu,n_i}(t_\mu)$ as a random variable indicating the number of incoming packets per mini-slot duration and $\eta^i_{m_\mu,n_i}(t_\mu)$ as the instantaneous packet size of URLLC UE $n_i \in \mathcal{N}_i = \{1, 2, \ldots, N_i\}$ belonging to the class i in the mini-slot m_μ of time slot t_μ. Hence, the total incoming URLLC traffic in the time slot t_μ is equal to $D_{total}(t_\mu) = \sum_{m_\mu=1}^{M_\mu}\sum_{i=1}^{I}\sum_{n_i=1}^{N_i}\eta^i_{m_\mu,n_i}(t_\mu)D^i_{m_\mu,n_i}(t_\mu)$. As a result, the $\gamma^{eMBB}_{k_\mu}(t_\mu)$ can be formulated as follows:

$$\begin{aligned}\gamma^{eMBB}_{k_\mu}(t_\mu) &= \sum_{b_\mu=1}^{B_\mu}\gamma^{eMBB}_{k_\mu b_\mu}(t_\mu) \\ &= \left[\sum_{b_\mu=1}^{B_\mu} x_{k_\mu b_\mu}(t_\mu) f_{b_\mu} \times \rho_{k_\mu b_\mu}(t_\mu)\frac{D_{total}(t_\mu)}{|B_\mu| \times |M_\mu|}\right]\end{aligned} \tag{3}$$

where $\rho_{k_\mu b_\mu}(t_\mu) \in [0, 1]$ indicates the weight of puncturing, and $|B_\mu| \times |M_\mu|$ presents the total system capacity in terms of frequency-time resources. The URLLC traffic is upper bounded by total system capacity, i.e., $D_{total}(t_\mu) \leq |B_\mu| \times |M_\mu|$. The $\rho_{k_\mu b_\mu}(t_\mu)$ identifies the pattern of overlapping total URLLC traffic in the time slot t_μ on the eMBB user k_μ resources in order to utilize (puncture) them for the URLLC transmission. The loss function is bounded $\gamma^{eMBB}_{k_\mu}(t_\mu) \in [0, \sum_{b_\mu=1}^{B_\mu} x_{k_\mu b_\mu}(t_\mu) f_{b_\mu}]$. Hence, for each eMBB user k_μ in time slot t_μ if:

- $\gamma^{eMBB}_{k_\mu}(t_\mu) = 0$, no puncturing;
- $0 < \gamma^{eMBB}_{k_\mu}(t_\mu) < \sum_{b_\mu=1}^{B_\mu} x_{k_\mu b_\mu}(t_\mu) f_{b_\mu}$, partial puncturing;
- $\gamma^{eMBB}_{k_\mu}(t_\mu) = \sum_{b_\mu=1}^{B_\mu} x_{k_\mu b_\mu}(t_\mu) f_{b_\mu}$, full puncturing happens.

Regarding URLLC traffic, the data rate of URLLC UE $n_i \in \mathcal{N}_i = \{1, 2, \ldots, N_i\}$ belonging to the class i can be approximated as [51], [52]:

$$r^{URLLC}_{n_{i\mu}}(t_\mu) = \sum_{k_\mu=1}^{K_\mu}\left[\left(\frac{\gamma^{eMBB}_{k_\mu}(t_\mu)}{\sum_{i=1}^{I} N_i} \times \log_2\left(1 + \frac{\sum_{j=1}^{J} p^j_{n_{i\mu}b_\mu}(t_\mu) h^j_{n_{i\mu}b_\mu}(t_\mu)}{\sigma^2_{Total_{n_{i\mu}}}}\right)\right) - \sum_{b_\mu=1}^{B_\mu}\Psi^{URLLC}_{n_{i\mu}b_\mu}(t_\mu)\right] \tag{4}$$

where $p^j_{n_{i\mu}b_\mu}(t_\mu)$ is the transmission power from the antenna j of the gNB over the RB b_μ to the URLLC user n_i in time slot t_μ; $h^j_{n_{i\mu}b_\mu}(t_\mu)$ is the Rayleigh fading channel gain of the transmission from the antenna j of the gNB over the RB b_μ to the URLLC user n_i in time slot t_μ; and $\sigma^2_{Total_{n_{i\mu}}}$ specifies total interference and noise power which negatively affects the URLLC user n_i in numerology μ. The $\Psi^{URLLC}_{n_{i\mu}b_\mu}(t_\mu)$ indicates the finite block-length channel coding regime in order to calculate the achievable rate of URLLC users which is given as:

$$\begin{aligned}&\Psi^{URLLC}_{n_{i\mu}b_\mu}(t_\mu) \\ &= \sqrt{\frac{1 - \left(1 + \frac{p^j_{n_{i\mu}b_\mu}(t_\mu) h^j_{n_{i\mu}b_\mu}(t_\mu)}{\sigma^2_{Total_{n_{i\mu}}}}\right)^{-2}}{C^{URLLC}_{n_{i\mu}b_\mu}(t_\mu)}} \times \frac{Q^{-1}(\varepsilon^d_{n_i})}{ln(2)}\end{aligned} \tag{5}$$

where $C^{URLLC}_{n_{i\mu}b_\mu}$ is the number of symbols in the mini-slot m_μ of time slot t_μ for the URLLC user n_i over the RB b_μ; and $Q^{-1}(.)$ is the inverse of the Gaussian Q-function of the decoding error probability for the URLLC user n_i.

The latency requirement of the URLLC user n_i over a particular numerology μ needs to satisfy the following [10], [21]:

$$\sum_{m_\mu=1}^{M_\mu}\eta^i_{m_\mu,n_i}(t_\mu).D^i_{m_\mu,n_i}(t_\mu) \leq r^{URLLC}_{n_{i\mu}}(t_\mu) \tag{6}$$

which indicates that the achieved data rate of URLLC user n_i has to be greater than the total incoming URLLC traffic of this particular URLLC user n_i in the time slot t_μ in order to satisfy its latency requirement in the numerology μ.

Regarding the reliability condition, we should know that the requests from all the URLLC users n_i of all the classes I within time slot t_μ have to be served in order to ensure that the reliability is satisfied. Thus, if θ^i_{max} ($\theta^i_{max} \ll 1$) represents

the outage probability threshold of the URLLC class i, we can define the reliability for each class as follows [9]:

$$Pr\left[\sum_{n_i=1}^{N_i} r_{n_{i\mu}}^{URLLC}(t_\mu) \le \sum_{n_i=1}^{N_i}\sum_{m_\mu=1}^{M_\mu} \eta_{m_\mu,n_i}^{i}(t_\mu).D_{m_\mu,n_i}^{i}(t_\mu)\right] \le \theta_{max}^{i} \tag{7}$$

which shows that the probability, in which the total number of served URLLC users (N_i) is less than the incoming URLLC traffic of all the users within URLLC class i, has to be less than θ_{max}^{i} in order to satisfy the reliability requirement.

As a result, the objective of the optimal resource allocation problem is to maximize the sum data rate of eMBB users over all utilized numerologies while ensuring that the individual data rate of each eMBB user is lower bounded by the minimum acceptable eMBB data rate, i.e., R_{min} to guarantee the fairness between eMBB users. Besides, at the same time, the resource allocation problem is required to fulfill the URLLC UEs' requirements in terms of extra low latency and ultra-high reliability. Consequently, the sum data rate maximization problem is formulated as follows:

$$\boldsymbol{P_0}: \max_{x,p,\rho} \ \mathbb{E}\left\{\sum_{\mu=0}^{4}\sum_{k_\mu=1}^{K_\mu}\sum_{t_\mu=0}^{T_\mu} R_{k_\mu}^{eMBB}(t_\mu)\right\} \tag{8a}$$

subject to :

$$\sum_{\mu=0}^{4}\sum_{k_\mu=1}^{K_\mu}\sum_{b_\mu=1}^{B_\mu}\sum_{j=1}^{J} p_{k_\mu b_\mu}^{j}(t_\mu) + \sum_{\mu=0}^{4}\sum_{i=1}^{I}\sum_{n_i=1}^{N_i}\sum_{b_\mu=1}^{B_\mu}\sum_{j=1}^{J} p_{n_{i\mu}b_\mu}^{j}(t_\mu) \le P_{max},$$

$$p_{k_\mu b_\mu}^{j}(t_\mu) \ge 0,\ p_{n_{i\mu}b_\mu}^{j}(t_\mu) \ge 0 \tag{8b}$$

$$\sum_{\mu=0}^{4}\sum_{k_\mu=1}^{K_\mu} x_{k_\mu b_\mu}(t_\mu) \le 1, \qquad x_{k_\mu b_\mu}(t_\mu) \in \{0,1\},$$

$$\forall b_\mu \in \mathcal{B}_\mu, \quad \forall k_\mu \in \mathcal{K}_\mu, \quad \forall \mu \in \boldsymbol{\mu} \tag{8c}$$

$$\sum_{k_\mu=1}^{K_\mu}\sum_{b_\mu=1}^{B_\mu} x_{k_\mu b_\mu}(t_\mu) \le |B_\mu|, \forall \mu \in \boldsymbol{\mu} \tag{8d}$$

$$\rho_{k_\mu b_\mu}(t_\mu) \in [0,1]$$

$$\forall b_\mu \in \mathcal{B}_\mu, \quad \forall k_\mu \in \mathcal{K}_\mu, \quad \forall \mu \in \boldsymbol{\mu} \tag{8e}$$

$$\sum_{m_\mu=1}^{M_\mu} \eta_{m_\mu,n_i}^{i}(t_\mu).D_{m_\mu,n_i}^{i}(t_\mu) \le r_{n_{i\mu}}^{URLLC}(t_\mu);$$

$$\forall n_i \in \mathcal{N}_i, \quad \forall i \in \mathcal{I}, \quad \forall \mu \in \boldsymbol{\mu} \tag{8f}$$

$$Pr\left[\sum_{n_i=1}^{N_i} r_{n_{i\mu}}^{URLLC}(t_\mu) \le \sum_{n_i=1}^{N_i}\sum_{m_\mu=1}^{M_\mu} \eta_{m_\mu,n_i}^{i}(t_\mu).D_{m_\mu,n_i}^{i}(t_\mu)\right] \le \theta_{max}^{i},$$

$$\forall i \in \mathcal{I}, \quad \forall \mu \in \boldsymbol{\mu} \tag{8g}$$

In this resource allocation problem, constraint (8b) defines the maximum transmission power budget via the gNB antennas in the downlink, P_{max}, towards all the eMBB and URLLC users. RBs' allocation among eMBB users is presented via constraint (8c). RBs' restriction in each numerology is presented by constraint (8d). Constraint (8e) indicates the weight of eMBB puncturing by the overlapped URLLC traffic. Finally, the latency and reliability requirements of the URLLC users are presented via (8f) and (8g), sequentially.

IV. OPTIMIZATION METHOD

In order to discover an optimal solution to the problem $\boldsymbol{P_0}$, it is necessary to find suitable mini-slots to position URLLC traffic on them while considering all potential RBs and power budget combinations of the eMBB users within different numerologies. Such a solution requires to satisfy eMBB users in terms of high data rate and, at the same time, URLLC users in terms of ultra-high reliability and extra low latency. Nonetheless, this procedure makes the solving approach complex as $\boldsymbol{P_0}$ is a non-convex problem. Since the optimization problem is mixed-integer nonlinear programming, we need to simplify this problem in order to reduce its complexity and make it solvable in a reasonable time. Hence, to find an appropriate solution to the $\boldsymbol{P_0}$ problem, we employ the decomposition and relaxation-based strategy for the eMBB and URLLC resource allocation problem. This results in converting $\boldsymbol{P_0}$ to a convex optimization problem. In this method, first, we decompose $\boldsymbol{P_0}$ into three sub-problems: $\boldsymbol{P_1}$ refers to the eMBB RBs allocation, $\boldsymbol{P_2}$ leads to the power allocation, and $\boldsymbol{P_3}$ considers URLLC traffic scheduling. Then, we relax the binary variable $x_{k_\mu b_\mu}(t_\mu)$ to a continuous variable in problem $\bar{\boldsymbol{P}}_{\boldsymbol{1}}$. Then, the fractional solution is rounded to get a solution to the original integer problem, $\boldsymbol{P_1}$. Subsequently, we also utilize Markov's inequality expression in order to linearly estimate (8g) requirement. Finally, we prove the convexity of $\bar{\boldsymbol{P}}_{\boldsymbol{1}}$, $\boldsymbol{P_2}$, and $\boldsymbol{P_3}$ sub-problems. Then each problem is solved individually based on its structure in order to achieve a practical solution with low computation complexity. The CVX toolbox [53], [54] is then used when solving each sub-problem.

A. EMBB RESOURCE BLOCK ALLOCATION PROBLEM

By decomposing $\boldsymbol{P_0}$ problem, while assuming p and ρ are constant values, the resource allocation problem, $\boldsymbol{P_1}$, is represented as follows:

$$\boldsymbol{P_1}: \max_{x} \ \mathbb{E}\left\{\sum_{\mu=0}^{4}\sum_{k_\mu=1}^{K_\mu}\sum_{t_\mu=0}^{T_\mu} R_{k_\mu}^{eMBB}(t_\mu)\right\} \tag{9a}$$

subject to :

$$\sum_{\mu=0}^{4}\sum_{k_\mu=1}^{K_\mu} x_{k_\mu b_\mu}(t_\mu) \le 1, \quad x_{k_\mu b_\mu}(t_\mu) \in \{0,1\},$$

$$\forall b_\mu \in \mathcal{B}_\mu, \quad \forall k_\mu \in \mathcal{K}_\mu, \quad \forall \mu \in \boldsymbol{\mu} \tag{9b}$$

$$\sum_{k_\mu=1}^{K_\mu}\sum_{b_\mu=1}^{B_\mu} x_{k_\mu b_\mu}(t_\mu) \le |B_\mu|, \forall \mu \in \boldsymbol{\mu} \tag{9c}$$

The existence of an integer variable in problem (9a) leads us to relax the $x_{k_\mu b_\mu}(t_\mu)$ to a continuous variable, $\bar{x}_{k_\mu b_\mu}(t_\mu)$, in order to avoid complexity in solving this problem. Hence, we can convert the original problem to $\bar{\boldsymbol{P}}_{\mathbf{1}}$ as follows:

$$\bar{\boldsymbol{P}}_{\mathbf{1}}: \max_{\bar{x}} \ \mathbb{E}\left\{\sum_{\mu=0}^{4}\sum_{k_\mu=1}^{K_\mu}\sum_{t_\mu=0}^{T_\mu} R_{k_\mu}^{eMBB}(t_\mu)\right\} + \upsilon\omega \tag{10a}$$

subject to :

$$\sum_{\mu=0}^{4}\sum_{k_\mu=1}^{K_\mu}\bar{x}_{k_\mu b_\mu}(t_\mu) \le 1+\omega \qquad 0 \le \bar{x}_{k_\mu b_\mu}(t_\mu) \le 1,$$
$$\forall b_\mu \in \mathcal{B}_\mu, \quad \forall k_\mu \in \mathcal{K}_\mu, \quad \forall \mu \in \boldsymbol{\mu} \tag{10b}$$

$$\sum_{k_\mu=1}^{K_\mu}\sum_{b_\mu=1}^{B_\mu}\bar{x}_{k_\mu b_\mu}(t_\mu) \le |B_\mu| + \omega, \forall \mu \in \boldsymbol{\mu} \tag{10c}$$

where $\omega = \max\{0, \sum_{k_\mu=1}^{K_\mu} x_{k_\mu b_\mu}(t_\mu) - 1\}$ is the rounding error value introduced by relaxing the integer variable, and υ is the weighting factor of ω with a negative value. The feasible solution to $\bar{\boldsymbol{P}}_{\mathbf{1}}$ is obtained with the minimum rounding error constraint, i.e., $\omega \to 0$.

Lemma 1: For constant values of p and ρ, $\bar{\boldsymbol{P}}_{\mathbf{1}}$ is a convex optimization problem.

Proof: It is worth noting that $R_{k_\mu}^{eMBB}(t_\mu)$ and its constraints are linear functions with respect to $\bar{x}_{k_\mu b_\mu}(t_\mu)$. The same applies to (10a) and its constraints, (10b) and (10c) concerning $\bar{x}_{k_\mu b_\mu}(t_\mu)$; thus, $\bar{\boldsymbol{P}}_{\mathbf{1}}$ is a convex optimization problem.

Finally, we need to convert the relaxed $\bar{x}_{k_\mu b_\mu}(t_\mu)$ variable back to the original binary variable $x_{k_\mu b_\mu}(t_\mu)$ after solving problem (9a). By determining $\alpha \in [0, 1]$ defined in [55], the conversion can be represented as:

$$x_{k_\mu b_\mu}(t_\mu) = \begin{cases} 1, & \text{if } \bar{x}_{k_\mu b_\mu}(t_\mu) \ge \alpha; \\ 0, & O.W. \end{cases} \tag{11}$$

■

B. EMBB POWER ALLOCATION PROBLEM

By decomposing $\boldsymbol{P_0}$ problem and presuming $\bar{x}$ and ρ as fixed values, the power allocation problem $\boldsymbol{P_2}$ is considered as follows:

$$\boldsymbol{P_2}: \max_{p} \ \mathbb{E}\left\{\sum_{\mu=0}^{4}\sum_{k_\mu=1}^{K_\mu}\sum_{t_\mu=0}^{T_\mu} R_{k_\mu}^{eMBB}(t_\mu)\right\} \tag{12a}$$

subject to :

$$\sum_{\mu=0}^{4}\sum_{k_\mu=1}^{K_\mu}\sum_{b_\mu=1}^{B_\mu}\sum_{j=1}^{J} p_{k_\mu b_\mu}^{j}(t_\mu) + \sum_{\mu=0}^{4}\sum_{i=1}^{I}\sum_{n_i=1}^{N_i}\sum_{b_\mu=1}^{B_\mu}\sum_{j=1}^{J} p_{n_{i\mu} b_\mu}^{j}(t_\mu) \le P_{max},$$
$$p_{k_\mu b_\mu}^{j}(t_\mu) \ge 0, \ p_{n_{i\mu} b_\mu}^{j}(t_\mu) \ge 0 \tag{12b}$$

Lemma 2: For fixed values of $\bar{x}$ and ρ, $\boldsymbol{P_2}$ is a convex optimization problem.

Proof: We calculate the Hessian matrix of $R_{k_\mu}^{eMBB}$ in order to investigate whether it is a convex or a concave function. According to the definition of a semi-definite matrix, we need to calculate the result of $z^T \times \boldsymbol{H_R} \times z$, which is a real number. In this expression, z is a real column vector, z^T is the transpose of z, and $\boldsymbol{H_R}$ is the Hessian matrix of $R_{k_\mu}^{eMBB}$ which is defined as follows:

$$\boldsymbol{H_R} = \begin{bmatrix} \frac{\partial^2 R}{\partial \bar{x}^2} & \frac{\partial^2 R}{\partial \bar{x}\partial \rho} & \frac{\partial^2 R}{\partial \bar{x}\partial p} \\ \frac{\partial^2 R}{\partial \rho \partial \bar{x}} & \frac{\partial^2 R}{\partial \rho^2} & \frac{\partial^2 R}{\partial \rho \partial p} \\ \frac{\partial^2 R}{\partial p \partial \bar{x}} & \frac{\partial^2 R}{\partial p \partial \rho} & \frac{\partial^2 R}{\partial p^2} \end{bmatrix} \tag{13}$$

Since in $\boldsymbol{P_2}$ we consider the $\bar{x}$ and ρ as fixed values in $R_{k_\mu}^{eMBB}$, thus, all of the $\boldsymbol{H_R}$ elements except the $(\boldsymbol{H_R})_{3,3} = \frac{\partial^2 R}{\partial p^2}$ are zero. By taking the second-order derivative of the $R_{k_\mu}^{eMBB}$ with respect to $p_{k_\mu b_\mu}^{j}(t_\mu)$ we obtain:

$$(\boldsymbol{H_R})_{3,3} = \frac{\partial^2 R}{\partial p^2} = -\frac{\sum_{b_\mu=1}^{B_\mu}\left[\left(x_{k_\mu b_\mu}(t_\mu) f_{b_\mu} - \gamma_{k_\mu b_\mu}^{eMBB}(t_\mu)\right) \times \left(\frac{\sum_{j=1}^{J} h_{k_\mu b_\mu}^{j}(t_\mu)}{\sigma_{Total_{k_\mu}}^2}\right)^2\right]}{ln(2) \times \left(1 + \frac{\sum_{j=1}^{J} p_{k_\mu b_\mu}^{j}(t_\mu) h_{k_\mu b_\mu}^{j}(t_\mu)}{\sigma_{Total_{k_\mu}}^2}\right)^2} \tag{14}$$

which obviously is always negative for any $p_{k_\mu b_\mu}^{j}(t_\mu)$ value. Now we calculate the result of $z^T \boldsymbol{H_R} z$ as follows:

$$z = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} \quad \forall z \in \mathbb{R}^3 \tag{15}$$

$$z^T \times \boldsymbol{H_R} \times z = z_3 (\boldsymbol{H_R})_{3,3} z_3 = (\boldsymbol{H_R})_{3,3} {z_3}^2 \le 0 \tag{16}$$

The result indicates that $\boldsymbol{H_R}$ is a negative semi-definite matrix, and consequently, $R_{k_\mu}^{eMBB}$ is a concave function. Since we want to maximize $\boldsymbol{P_2}$ and due to the linearity constraint of (12b) with respect to $p_{k_\mu b_\mu}^{j}(t_\mu)$, $\boldsymbol{P_2}$ is a convex optimization problem. ■

C. URLLC TRAFFIC SCHEDULING

In this section, first, we employ the Markov inequality expression [56] in order to simplify the constraint (8g) to a linear condition as follows:

$$Pr\left[\sum_{n_i=1}^{N_i} r_{n_{i\mu}}^{URLLC}(t_\mu) \le \sum_{n_i=1}^{N_i}\sum_{m_\mu=1}^{M_\mu} \eta_{m_\mu,n_i}^{i}(t_\mu).D_{m_\mu,n_i}^{i}(t_\mu)\right]$$
$$\le \frac{\mathbb{E}\left[\sum_{n_i=1}^{N_i}\sum_{m_\mu=1}^{M_\mu} \eta_{m_\mu,n_i}^{i}(t_\mu).D_{m_\mu,n_i}^{i}(t_\mu)\right]}{\sum_{n_i=1}^{N_i} r_{n_{i\mu}}^{URLLC}(t_\mu)} \le \theta_{max}^{i} \tag{17}$$

Finally, decomposing $\boldsymbol{P_0}$ problem, while supposing $\bar{x}$ and p as invariant, yields in URLLC scheduling problem, $\boldsymbol{P_3}$, as follows:

$$\boldsymbol{P_3}: \max_{\rho} \ \mathbb{E}\left\{\sum_{\mu=0}^{4}\sum_{k_\mu=1}^{K_\mu}\sum_{t_\mu=0}^{T_\mu} R_{k_\mu}^{eMBB}(t_\mu)\right\} \tag{18a}$$

subject to :

$$\rho_{k_\mu b_\mu}(t_\mu) \in [0,1]$$
$$\forall b_\mu \in \mathcal{B}_\mu, \quad \forall k_\mu \in \mathcal{K}_\mu, \quad \forall \mu \in \boldsymbol{\mu} \tag{18b}$$

$$\sum_{m_\mu=1}^{M_\mu} \eta^i_{m_\mu,n_i}(t_\mu).D^i_{m_\mu,n_i}(t_\mu) \le r^{URLLC}_{n_{i\mu}}(t_\mu);$$
$$\forall n_i \in \mathcal{N}_i, \quad \forall i \in \mathcal{I} \tag{18c}$$

$$\frac{\mathbb{E}\left[\sum_{n_i=1}^{N_i}\sum_{m_\mu=1}^{M_\mu} \eta^i_{m_\mu,n_i}(t_\mu).D^i_{m_\mu,n_i}(t_\mu)\right]}{\theta^i_{max}} \le \sum_{n_i=1}^{N_i} r^{URLLC}_{n_{i\mu}}(t_\mu),$$
$$\forall n_i \in \mathcal{N}_i, \quad \forall i \in \mathcal{I} \tag{18d}$$

Lemma 3: For invariant values of $\bar{x}$ and p, $\boldsymbol{P_3}$ is a convex optimization problem.

Proof: By considering the equivalent value of the loss function expressed in (3) and substituting it in (2), it is evident that $R^{eMBB}_{k_\mu}(t_\mu)$ is a linear function with respect to $\rho_{k_\mu b_\mu}(t_\mu)$. Since (18c) and (18d) are also linear constraints with respect to $\rho_{k_\mu b_\mu}(t_\mu)$; thus, $\boldsymbol{P_3}$ is a convex optimization problem. ■

D. MIMRA AS A SOLUTION OF PROBLEM (8a)

In this section, we present our proposed *Mixed-numerology Mini-slot based Resource Allocation (MiMRA)* algorithm to find an optimal solution for Eq. (8a). First, MiMRA algorithm converts $\boldsymbol{P_0}$ to $\bar{\boldsymbol{P_1}}$, $\boldsymbol{P_2}$, and $\boldsymbol{P_3}$ sub-problems by relaxation and decomposition technique. Next, the algorithm sets the minimum acceptable eMBB data rate, i.e., R_{min}, in order to guarantee fairness between eMBB users. In this way, the algorithm not only maximizes the sum rate of the eMBB users but also ensures that each individual eMBB user will achieve at least the R_{min} value. In each iteration l the algorithm searches for $\bar{\boldsymbol{x}}^{(l)}, \boldsymbol{p}^{(l)}, \boldsymbol{\rho}^{(l)}$ as a solution from a feasible convex set. Then within each numerology μ, the algorithm specifies the number of punctured mini-slots for a particular eMBB user in that numerology, k_μ, as follows:

$$\pi_{k_\mu} = \left\lceil \frac{\rho_{k_\mu b_\mu}(t_\mu) D_{total}(t_\mu)}{|B_\mu|} \right\rceil \tag{19}$$

where $\pi_{k_\mu} \in \{0, 1, 2, \ldots, M_\mu\}$. The MiMRA algorithm also defines a *puncturing rate threshold*, i.e., $th^{eMBB}(t_\mu)$, according to the loss functions for all eMBB users. The selection criteria for calculating $th^{eMBB}(t_\mu)$ is as follows:

$$th^{eMBB}(t_\mu) = \begin{cases} \max_{\forall k_\mu \in \mathcal{K}_\mu}\left\{\gamma^{eMBB}_{k_\mu}(t_\mu)\right\}, \\ 0 \le \gamma^{eMBB}_{k_\mu}(t_\mu) < \sum_{b_\mu=1}^{B_\mu} x_{k_\mu b_\mu}(t_\mu) f_{b_\mu}; \\ \max_{\forall k_\mu \in \mathcal{K}_\mu}\left\{\gamma^{eMBB}_{k_\mu}(t_\mu)\right\} - offset_\mu, \\ \gamma^{eMBB}_{k_\mu}(t_\mu) = \sum_{b_\mu=1}^{B_\mu} x_{k_\mu b_\mu}(t_\mu) f_{b_\mu}; \end{cases} \tag{20}$$

where $offset_\mu$ indicates a constant value to tune $th^{eMBB}(t_\mu)$ if the second condition in (20) holds. After setting a value for $th^{eMBB}(t_\mu)$, the algorithm proceeds to calculate the number of punctured mini-slots for each eMBB user and the achievable data rate to verify whether each eMBB user can at least attain R_{min} or not. Thereafter, if $\sum_{\mu=0}^{4}\sum_{t_\mu=0}^{T_\mu} R^{eMBB}_{k_\mu}(t_\mu) < R_{min}$, then, depending on the $\rho_{k_\mu b_\mu}(t_\mu) \in [0,1]$, the algorithm maps part or the whole of incoming URLLC load, $D_{total}(t_\mu)$, to another possible eMBB user k'.

gNB holds the channel state information of the users. It continuously tracks the eMBB users within its coverage area by keeping a log of their channel conditions and the distance they are located. gNB searches to find the potential eMBB user k' with the allocated RB in the same numerology, b'_μ, or with RB in another numerology, $b'_{\mu'}$, if at least one of the following six conditions in Eq. (21) is fulfilled. Otherwise the algorithm runs for another round of iteration, $l+1$, to find another possible set of solutions, i.e., $\bar{\boldsymbol{x}}^{(l+1)}, \boldsymbol{p}^{(l+1)}, \boldsymbol{\rho}^{(l+1)}$ until it converges.

$$\begin{cases} \text{a) if } k' \text{ is allocated a higher power than } k \text{ in } \mu \text{ or } \mu': \\ \begin{cases} p^{j'}_{k'_\mu b'_\mu}(t_\mu) > p^{j}_{k_\mu b_\mu}(t_\mu); & \text{in } \mu \\ p^{j'}_{k'_{\mu'} b'_{\mu'}}(t_{\mu'}) > p^{j}_{k_\mu b_\mu}(t_\mu); & \text{in } \mu' \end{cases} \\ \text{b) if } k' \text{ has a larger channel gain than } k \text{ in } \mu \text{ or } \mu': \\ \begin{cases} h^{j'}_{k'_\mu b'_\mu}(t_\mu) > h^{j}_{k_\mu b_\mu}(t_\mu); & \text{in } \mu \\ h^{j'}_{k'_{\mu'} b'_{\mu'}}(t_{\mu'}) > h^{j}_{k_\mu b_\mu}(t_\mu); & \text{in } \mu' \end{cases} \quad \mathbf{k} \xrightarrow{D_{total}(t_\mu)} \mathbf{k}' \\ \text{c) if } k' \text{ has a lower loss function than } k \text{ in } \mu \text{ or } \mu': \\ \begin{cases} \gamma^{eMBB}_{k'_\mu b'_\mu}(t_\mu) < \gamma^{eMBB}_{k_\mu b_\mu}(t_\mu); & \text{in } \mu \\ \gamma^{eMBB}_{k'_{\mu'} b'_{\mu'}}(t_{\mu'}) < \gamma^{eMBB}_{k_\mu b_\mu}(t_\mu); & \text{in } \mu' \end{cases} \end{cases} \tag{21}$$

As a result, the proposed algorithm tries to maximize the sum rate of all eMBB users while also considering each individual eMBB user to achieve the minimum acceptable data rate to fulfill the QoS requirement. Algorithm 1 summarizes the above steps.

E. COMPLEXITY ANALYSIS OF THE PROPOSED ALGORITHM

This subsection represents the computational complexity of the proposed MiMRA algorithm. In order to calculate the complexity of the algorithm, we notice that it is composed of some nested loops. The computational complexity of MiMRA is $\mathcal{O}(|L||\boldsymbol{\mu}||I||T_\mu||K_\mu||J|)$. Nevertheless, the $|\boldsymbol{\mu}|$, $|I|$, and $|J|$ have finite values and cannot get very high arbitrary values. As we know, the largest value of numerologies, $|\boldsymbol{\mu}|$, is 4. In addition, according to [57], the URLLC traffic classes, $|I|$, are mainly categorized into up to 8 different classes. Finally, according to [58], the maximum number of antennas that so far have been practically implemented in Massive MIMO base stations is 64. Consequently

Algorithm 1 MiMRA Algorithm for eMBB/URLLC Coexistence

Input: $\mu \in \boldsymbol{\mu}, i \in I, t \in T, b \in \mathcal{B}, k \in \mathcal{K}$, $j \in \mathcal{J}, h^{j}_{kb}(t), P_{max}$
Output: Solution to Eq. (8a) and providing fairness between eMBB users.
Relax $\boldsymbol{x}$ to $\bar{\boldsymbol{x}}$, and decompose $\boldsymbol{P_0}$ to $\bar{\boldsymbol{P}}_1$, $\boldsymbol{P_2}$, and $\boldsymbol{P_3}$
Set R_{min}
for $l \leftarrow 0$ to L **do**
  Find $\bar{\boldsymbol{x}}^{(l)}, \boldsymbol{p}^{(l)}, \boldsymbol{\rho}^{(l)}$ from feasible convex set as a solution of $\bar{\boldsymbol{P}}_1$, $\boldsymbol{P_2}$, and $\boldsymbol{P_3}$ respectively.
  Find $\boldsymbol{x}^{(l)}$ via Eq. (11).
  Define $th^{eMBB}(t_\mu)$ according to Eq. (20).
  for $\mu \in \boldsymbol{\mu}$ **do**
    for $i \in I$ **do**
      for $t_\mu \in T_\mu$ **do**
        for $k_\mu \in K_\mu$ **do**
          for $j \in J$ **do**
            Calculate π_{k_μ}, $R^{eMBB}_{k_\mu}(t_\mu)$ based on $th^{eMBB}(t_\mu)$
            if $\sum_{\mu=0}^{4} \sum_{t_\mu=0}^{T_\mu} R^{eMBB}_{k_\mu}(t_\mu) < R_{min}$ **then**
              According to $\rho_{k_\mu b_\mu}(t_\mu)$ map part or the whole $D_{total}(t_\mu)$ to k' in case of Eq. (21).
            else
              Go back to step: (7).
            end if
          end for
        end for
      end for
    end for
  end for
end for

the actual total computational complexity of MiMRA is $\mathcal{O}(|L||T_\mu||K_\mu|)$.

V. PERFORMANCE EVALUATION

In this section, we demonstrate the efficiency of our proposed algorithm through simulations and evaluate the performance of the algorithm.

A. NETWORK SCENARIO

We consider a shared 5G NR infrastructure with several URLLC and eMMB users in coexistence as illustrated in Figure 2. Notice that in this scenario, we would like to serve the generated URLLC packets belonging to different URLLC classes according to their priority, defined in Table 5. The common RAN physical resources are logically shared to transmit URLLC and eMBB traffic towards corresponding users in the downlink. The URLLC traffic is generated by the power distribution and manufacturing verticals belonging to two distinct URLLC classes. The eMBB traffic is produced from video streaming of a popular sport tournament. The different types of traffic are characterized by the following scenario which is used to determine the simulation parameters.

- URLLC traffic class 1: Generated by the IEDs placed in power distribution grids which broadcast GOOSE messages when an event (e.g., alarm, failure, or any mission-critical event) occurs [59]. We imagine that a failure occurs in the observed geographical area (which falls under the coverage of the gNB) and investigate the impact of the injected GOOSE messages into the network.
- URLLC traffic class 2: We assume a large manufacturing factory continuously operating in the same geographical area. Few sensors are installed inside the processing section to obtain measurements and perform process automation.
- eMBB traffic: Meantime, a largely popular sport event is assumed to be happening thus, several residents in the area are video streaming the live broadcast of the event with the HD quality up to 4K resolution.

TABLE 5. Simulation parameter configurations.

Type of traffic	URLLC class $i = 1$	URLLC class $i + 1 = 2$	eMBB
No. of users	10	10	20
Air latency [1] [ms]	1 [2]	5	-
Minimum data rate [Mbps]	1.5 [3]	1	-
Priority	high	medium	low
Traffic model	Poisson	Poisson	Full-buffered

[1] It is considered that standard air latency requirements are assigned to 20% of the corresponding E2E latency requirements [57], [60].
[2] Fault case.
[3] Fault case.

B. SIMULATION SETUP

We study and simulate the 5G RAN domain in a dense urban microcell scenario. In our simulated 5G NR, we consider one gNB operating in the FR-1 with 8 antennas towards the downlink, located at the center of the cell coverage zone with a 500 m radius. The operating center frequency is set to 3.5 GHz and $P_{max} = 40$ dBm. Several single antenna eMBB and URLLC users are randomly distributed within the coverage zone. Besides, the gNB schedules eMBB and URLLC transmissions over flat i.i.d Rayleigh fading channels. The remaining system parameters are listed in Table 5. In order to provide practical results comparable to realistic scenarios, the target KPI values of eMBB and URLLC services are extracted from specification documents [60].

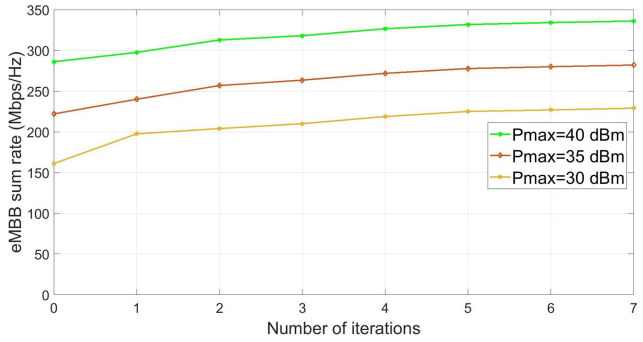


FIGURE 3. Convergence of MiMRA algorithm with various gNB maximum transmit power, $P_{max} = \{30, 35, 40\}$ dBm, and $R_{min} = 8$Mbps.

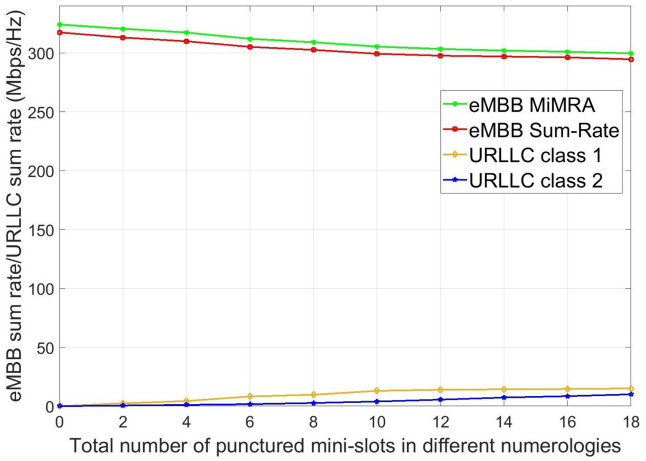


FIGURE 4. Achievable eMBB/URLLC sum rates over the total number of punctured mini-slots in different numerologies.

Extensive simulations are carried out under the following situations. It is considered that the gNB punctures the pre-scheduled eMBB traffic towards the corresponding users in the downlink by transmitting the overlapped URLLC traffic classes modeled as Poisson processes. In addition, air latency of the URLLC traffic classes is also considered in the simulation while assuming the eMBB traffic is not delay-sensitive compared to the URLLC traffic. The gNB utilizes numerologies $\mu = 0, 1, 2$ to transmit eMBB traffic over all of the available RBs in each numerology [40], [61]. The corresponding time slots for each numerology, $t_{\mu=0} = 1$ ms, $t_{\mu=1} = 0.5$ ms, and $t_{\mu=2} = 0.25$ ms, are subdivided into number of M_0, M_1, and M_2 mini-slots, respectively. The gNB punctures the eMBB traffic with URLLC traffic class $i = 1, i + 1 = 2$ over all the utilized numerologies. We first evaluate the performance of the network in terms of the achievable sum rate of the eMBB and different classes of URLLC users under the total number of punctured mini-slots in different numerologies. Then we investigate the sum rate of the eMBB users for two various minimum acceptable rates per eMBB user under the number of URLLC packets generated from several URLLC users within class $i = 1, i+1 = 2$. After that, we repeat the previous evaluation, but this time under different gNB transmit power values. Next, we analyze the obtainable sum rate of the eMBB users for two diverse maximum allowed delay requirements of the URLLC users under different gNB transmit power values.

C. PERFORMANCE RESULTS

First, we evaluate the convergence speed of the MiMRA algorithm. As illustrated in Figure 3, we investigate how fast MiMRA converges according to the different values for the maximum transmit power, P_{max}, of gNB towards different users. We can observe that the eMBB sum rate converges fast and evolves to saturated status after around 5 iterations. Besides, it can be noticed that we have a higher eMBB sum rate employing higher transmit power. In particular, the eMBB sum rate can reach up to 336Mbps for $P_{max} = 40$ dBm. In contrast, the eMBB sum rate can obtain less value for smaller transmit power from the gNB. It is obvious that the reason is because of having a higher SNR value for larger P_{max}, which results in holding a larger eMBB data rate.

In Figure 4, we illustrate the achievable sum rates of URLLC classes $i = 1, i + 1 = 2$ and eMBB users over the total number of punctured mini-slots in different numerologies. In this part, we set $R_{min} = 15$Mbps. The figure represents a gradual decrease/increase of the achievable eMBB/URLLC sum rates, respectively. In this figure, we compare the performance of the MiMRA algorithm with the baseline approach, *Sum-Rate* [62] scheduler, whose objective is to maximize the average sum rate of eMBB users via involving the puncturing strategy. As the figure reveals, the eMBB users achieve their maximum sum rate with no punctured mini-slots. By receiving URLLC traffic from either class, $i = 1$ or $i + 1 = 2$ or both, gNB starts puncturing the eMBB users over various numerologies depending on the number of RBs required to fulfill the URLLC users, the priority of the URLLC users, and their latency requirements. As a result, the gNB assigns the demanded RBs to the URLLC class $i = 1$ and then URLLC class $i+1 = 2$ due to their QoS requirements. As the gNB tries to satisfy the URLLC users, the sum rate of the eMBB users decreases. Specifically, by puncturing up to 18 mini-slots in various numerologies and assigning the necessary RBs, the gNB serves the URLLC users of class $i = 1$ and $i + 1 = 2$ to reach their sum rate up to 15Mbps and 10Mbps, respectively. These sum rate values are appropriate to transmit the URLLC packets towards the corresponding URLLC users in different classes in the downlink. Regarding the eMBB users, as it can be seen from the figure, the MiMRA algorithm outperforms the *Sum-Rate* since even with a high number of punctured mini-slots (18 mini-slots), the MiMRA algorithm is still able to deliver the minimum acceptable data rate for each eMBB user $R_{min} = 15$Mbps to provide of up to 299.57Mbps as the sum rate of the eMBB users.

We next evaluate the sum rates of the eMBB users according to the allocated power from the gNB. We consider two

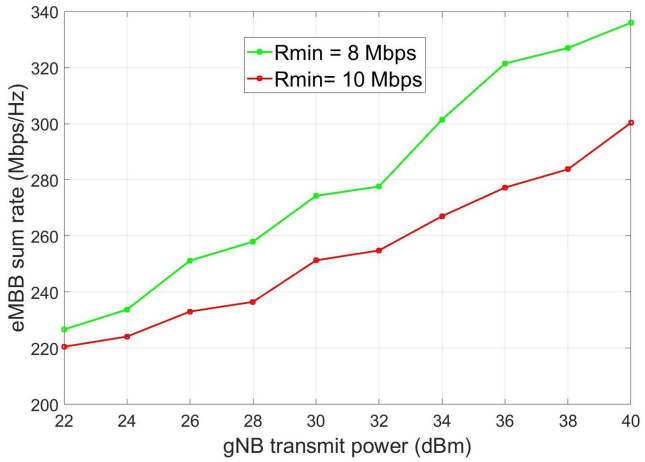


FIGURE 5. Achievable sum rates of eMBB users with two minimum acceptable data rate values versus different gNB transmit power.

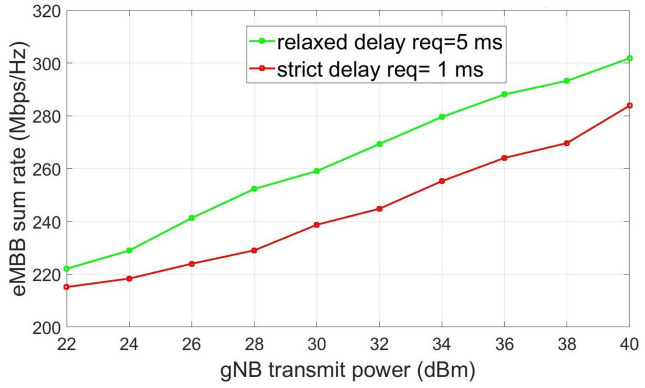


FIGURE 6. Achievable sum rates of eMBB users affected by traffic from two URLLC classes with two maximum allowed delay requirements versus different gNB transmit power.

scenarios. We set $R_{min} =$ 8Mbps and 10Mbps as the minimum acceptable data rates for the QoS requirements of the eMBB users in these scenarios. As Figure 5 illustrates, we can observe that higher QoS requirement for R_{min} results in lower sum rate performance due to stringent service requirements. The gNB potentially requires more resources to serve those highly-demanding eMBB users. Nevertheless, delivering the $R_{min} =$ 8Mbps for each eMBB user is more feasible and attainable rather than the $R_{min} =$ 10Mbps in case of increasing the number of eMBB and URLLC users.

Figure 6 demonstrates a follow-up graph to present the achievable sum rates of eMBB users impacted by traffic from two URLLC classes with two maximum delay requirements versus different gNB transmit power. As it can be observed, the sum rates of the eMBB users degrade with the stringent delay requirement. It means that in the case of strict latency, the system performance will confine more than the relaxed delay requirement case. The gNB requires more RBs with shorter time slots to satisfy the delay requirement of the URLLC class with stricter latency, 1ms, compared to the URLLC class with a relaxed latency, 5ms. Consequently, in

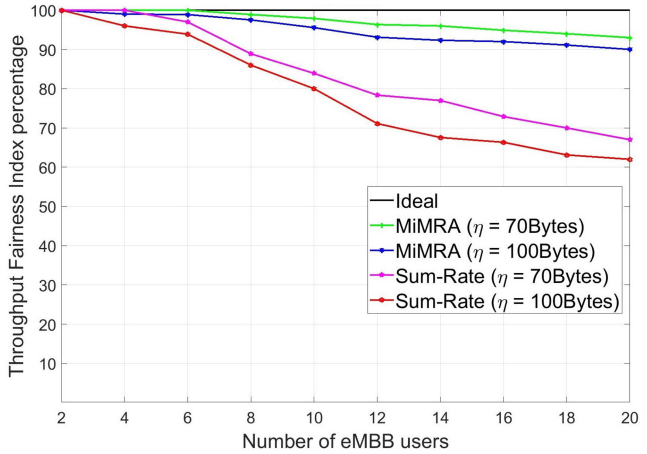


FIGURE 7. Fairness between eMBB users.

the case of incoming URLLC traffic with an even stricter value than 1 ms, the sum rates of the eMBB users are further reduced.

Figure 7 illustrates fairness in allocating the demanded resources between the eMBB users. We compare the performance of the MiMRA algorithm with the *Sum-Rate* scheduler under different packet sizes of the URLLC traffic classes. We want to investigate how much the fairness values provided by the MiMRA and *Sum-Rate* differ from the ideal (desired) case in which there is perfect fairness in allocating the required resources between the eMBB users. The fairness among the eMBB users is calculated based on Jain's Fairness index [63]. As it can be seen from the Figure, for a smaller packet size of $\eta =$ 70Bytes, the MiMRA algorithm performs well compared to the *Sum-Rate* outcome as MiMRA fairness is close to the ideal value (desired fairness) even for a large number of eMBB users (20 users). We observe the same performance for a larger packet size of $\eta =$ 100Bytes, as MiMRA again outperforms *Sum-Rate*. In both cases, there is a large difference between the fairness resulting from the MiMRA algorithm and the *Sum-Rate*. In particular, for $\eta =$ 70Bytes, MiMRA and *Sum-Rate* grant up to 93% and 67%, and for $\eta =$ 100Bytes, provide up to 90% and 62% fairness, respectively. This performance is due to Eq. (21) and considering $R_{min} =$ 8Mbps in order to find the perfect candidate for puncturing.

In Figure 8, we study the sum rates of the eMBB users versus the URLLC traffic load. In particular, we set $R_{min}=$ 6Mbps and 8Mbps as the two minimum acceptable data rates. Then, we evaluate the performance of the network for handling the incoming URLLC load with two maximum outage probability threshold values, $\theta^1_{max} = 0.001$ and $\theta^2_{max} = 0.01$, belonging to URLLC class $i = 1$ and $i + 1 = 2$, respectively. As observed from the Figure, the eMBB users can reach high values for their sum rates when the number of URLLC packets is zero. As the incoming URLLC traffic classes arrive with different outage probabilities, the gNB punctures eMBB users to serve the URLLC traffic types. In

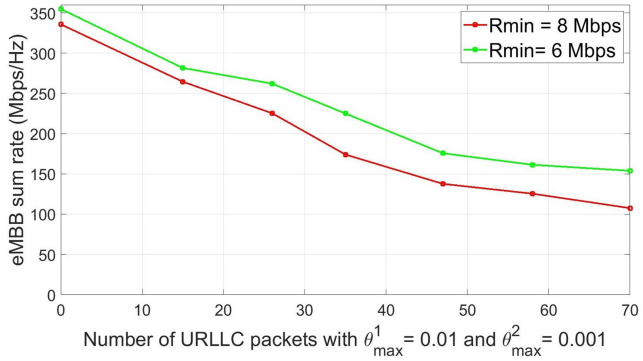


FIGURE 8. Achievable sum rates of eMBB users with two minimum acceptable data rate values URLLC load.

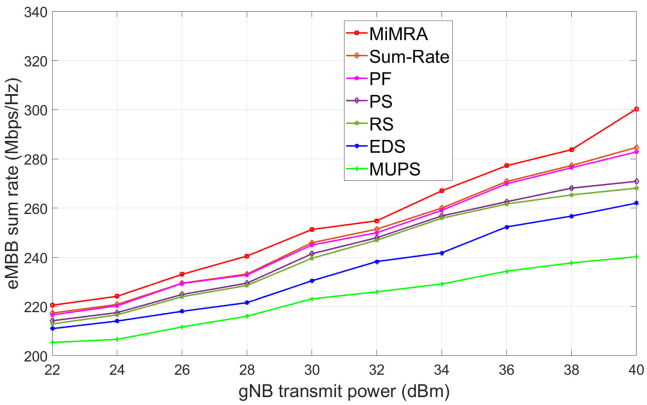


FIGURE 9. Comparison of eMBB sum rates with MiMRA and baselines for different gNB transmit power when R_{min} = 8Mbps.

such a situation, the gNB both tries to fulfill the requirements for two URLLC traffic classes, and to puncture eMBB users to the extent that each eMBB user can still achieve the minimum acceptable data rate. By considering the fairness provided by the MiMRA algorithm, for up to 38 URLLC packets, including both traffic classes, we can achieve the promised minimum data rate of up to 10.5Mbps per user if the R_{min} = 6Mbps. For the same number of URLLC packets, we reach up to 8Mbps per user if the R_{min} = 8Mbps. As the number of URLLC packets increases up to 70 packets, the gNB can still provide 7.65Mbps if the R_{min} = 6Mbps, but it ultimately can deliver up to utmost 5.37Mbps if the R_{min} = 8Mbps. It is worth mentioning that the MiMRA algorithm performs well since the sporadic behavior of URLLC traffic makes it rare to have such a high value of URLLC packets (70) in a concise period of a time slot.

Figure 9 illustrates the eMBB sum rates versus different gNB transmit power values. In particular, we compare the performance of MiMRA with 1) *Sum-Rate* that adopts a puncturing strategy to maximize the sum-rate of all eMBB users; 2) *Random Scheduler (RS)* [12] that transmits the incoming URLLC traffic by randomly picking pre-allocated RBs to the eMBB users; 3) *Proportional Fair (PF)* [64] which attempts to use the variations of channel conditions by

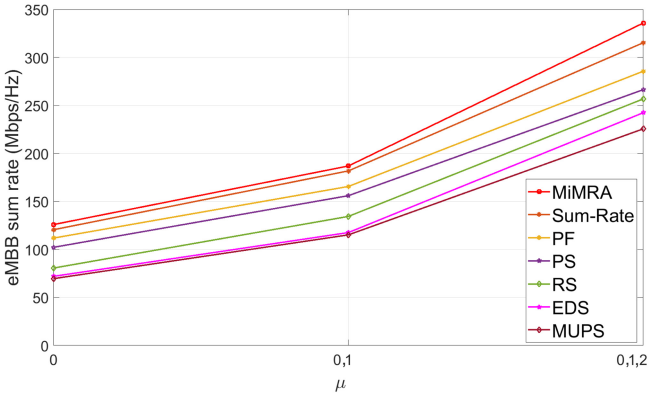


FIGURE 10. Comparison of MiMRA and baselines for eMBB sum rates versus different numerologies, μ, for R_{min} = 8Mbps, P_{max} = 40 dBm, and 10 URLLC packets.

assigning resources to users with the most suitable conditions for the upcoming period; 4) *Punctured Scheduling (PS)* [15] (also known as user-based puncturing) that chooses the RBs with the highest MCS assigned to eMBB users, and it punctures them to serve URLLC traffic; 5) *Equally Distributed Scheduler (EDS)* [10] which equally selects pre-allocated RBs to each of the eMBB users to serve the URLLC traffic; and 6) *Multi-User Preemptive Scheduling (MUPS)* [65]. We consider the incoming URLLC load θ^1_{max} = 0.001 and θ^2_{max} = 0.01, belonging to URLLC class i = 1 and $i+1$ = 2, respectively and in total 28 packets. It is observed from the Figure that for lower power values, P_{max} = 22 dBm, MiMRA provides up to 220.5Mbps. Under the same condition and URLLC load, *Sum-Rate*, *PF*, *PS*, *RS*, *EDS*, and *MUPS* grant up to 217.1Mbps, 216.4Mbps, 214.2Mbps, 212.2Mbps, 210.8Mbps, and 205Mbps respectively. This exposes that there is a 3.4Mbps gap between MiMRA and the second best algorithm, which is *Sum-Rate*. Besides, since MiMRA exhibits high fairness, it can be inferred that each eMBB user can achieve up to at least 11.2Mbps, which is still higher than the R_{min} = 8Mbps. MiMRA utilizes a higher power value, P_{max} = 40 dBm, to deliver up to 300Mbps in order to perform even better than before with the price of consuming higher power. In this case, *Sum-Rate*, *PF PS*, *RS*, *EDS*, and *MUPS* provide 282.1Mbps, 281Mbps, 269.8Mbps, 267.9Mbps, 261.3Mbps, and 240.1Mbps, respectively. The difference between MiMRA performance and *Sum-Rate* is almost 17.9Mbps. Considering the high fairness of MiMRA, it can deliver up to 15Mbps. As a result, by keeping the data rate per eMBB user higher than or close to the R_{min}, the network guarantees that each eMBB user receives at least minimum resources, which are required for full HD video streaming with very high resolution with almost zero buffer time. In fact, with the MiMRA algorithm, the gNB does not permit to puncture any of the eMBB users completely, and it maintains the data rate at a level to avoid decreasing per eMBB data rate per user remarkably.

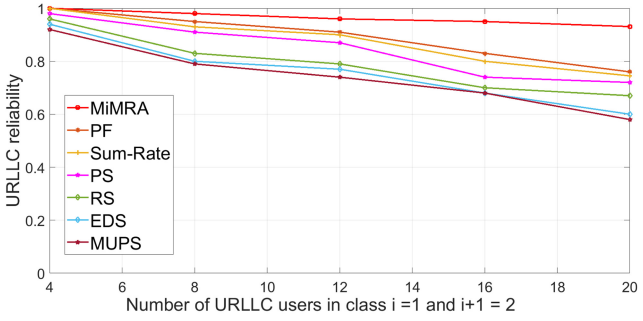
Figure 10 represents the performance of MiMRA in achieving a higher eMBB sum rate compared to baselines

with respect to the numerology values, μ. For $\mu = 0$, MiMRA shows a relatively similar value to what *Sum-Rate* delivers for the eMBB sum rate. However, MiMRA performance is slightly better than the *Sum-Rate*. As the value of μ increases, the eMBB sum rates of different approaches also grow. This is due to the increase in the SCS of RB with their respective numerologies. As the value of μ evolves, MiMRA outperforms the other solutions. In particular, MiMRA attains almost 336Mbps while *Sum-Rate* obtains 315Mbps over the employed numerologies.

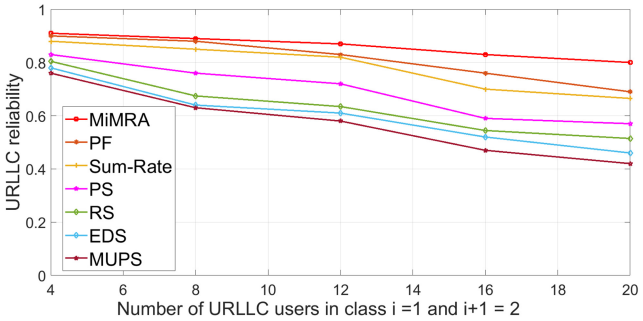
Usually in order to enhance *sum rate*, eMBB users with high channel gains need to be assigned more RBs than users with low channel gains. Nevertheless, this worsens the performance of eMBB users with poor channel conditions, especially if they are punctured by the incoming URLLC traffic as well. This results in acquiring significantly low fairness among the eMBB users. Hence, to maximize the *sum rate*, the number of RBs allocated to users with poor channel conditions has to be high. This creates a crucial dilemma between having high *sum rate* and fairness among the eMBB users [21]. As can be comprehended from the simulation results, the MiMRA algorithm resolves this challenge by setting R_{min} and defining $th^{eMBB}(t_\mu)$ to maximize the *sum rate* of the eMBB users and deliver fairness among them. Additionally, since the B5G era incorporates dealing with various URLLC use cases with distinct QoS requirements, MiMRA simultaneously ensures to fulfill diverse URLLC classes' demands for extra low latency and ultra-high reliability. Thus, URLLC traffic classes belonging to critical use cases are served with the highest priority, as discussed in the following.

Figure 11(a) and 11(b) illustrate the URLLC reliability of two classes (class 1 and class 2) with two packet sizes. As can be observed in 11(a), reliability drops as the number of URLLC users increases. In particular, with $\eta = 70$Bytes, for a few URLLC users (up to 12 users), MiMRA grants reliability of up to 96% while *PF* provides up to 90%. As the number of URLLC users grows (20 users), MiMRA still guarantees the URLLC reliability of up to 94% while *PF* can deliver maximum reliability of up to 76%. In the case of 11(b), we can see that the reliability reduces even further when the URLLC packet size increases. When the number of URLLC users is 20, MiMRA provides up to 80% reliability while the second best, *PF*, offers up to 69.8% reliability.

Figure 12 compares the URLLC delay Cumulative Distribution Function (CDF) of different baselines for 10 eMBB and 10 URLLC users of the mission-critical case with a delay requirement of 1ms. In particular, *MUPS* delivers the largest delay as it cannot counteract the strict URLLC delay requirement with an appropriate resource allocation that satisfies eMBB and URLLC users simultaneously. *EDS* provides the second largest delay, which still cannot satisfy the delay constraint of 1ms. Nevertheless, MiMRA outperforms other baseline solutions by meeting the delay requirement of URLLC packets.



(a) URLLC reliability for $\eta = 70$Bytes



(b) URLLC reliability for $\eta = 100$Bytes

FIGURE 11. Comparison of URLLC reliability with MiMRA and baselines versus different numbers of URLLC users with two URLLC packet sizes.

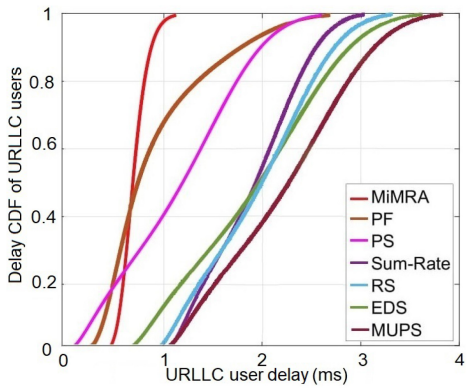


FIGURE 12. Delay CDF of the URLLC users in class $i = 1$ with a delay requirement of 1ms.

VI. CONCLUSION

In this paper, we propose an optimization framework to solve the resource allocation problem of coexisting URLLC and eMBB users in 5G NR over various numerologies. Furthermore, we study the impact of the incoming URLLC traffic, which is scheduled immediately into the mini-slots, instead of eMBB users, due to the stringent latency requirement. Our main goal is to maximize the sum rate for the eMBB users and to achieve a minimum acceptable data rate for the individual eMBB users ensuring fairness. The simulation results show that the proposed algorithm MiMRA enhances the sum rate of eMBB users while, at the same

time, each eMBB user can still achieve a minimum acceptable data rate. Thus, the eMBB users experience a more reliable transmission than the other studied approaches.

REFERENCES

[1] Q. Zhang and F. H. Fitzek, "Mission critical IoT communication in 5G," in *Proc. 1st Int. Conf. FABULOUS*, Ohrid, Republic of Macedonia, 2015, pp. 35–41.

[2] L. Bonati, M. Polese, S. D'Oro, S. Basagni, and T. Melodia, "Open, programmable, and virtualized 5G networks: State-of-the-art and the road ahead," *Comput. Netw.*, vol. 182, Dec. 2020, Art. no. 107516.

[3] D. Kreutz, F. M. Ramos, P. E. Verissimo, C. E. Rothenberg, S. Azodolmolky, and S. Uhlig, "Software-defined networking: A comprehensive survey," *Proc. IEEE*, vol. 103, no. 1, pp. 14–76, Jan. 2015.

[4] F. Z. Yousaf, M. Bredel, S. Schaller, and F. Schneider, "NFV and SDN—Key technology enablers for 5G networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 11, pp. 2468–2478, Nov. 2017.

[5] P. Costa, M. Migliavacca, P. Pietzuch, and A. L. Wolf, "NaaS: Network-as-a-service in the cloud," in *Proc. 2nd USENIX Workshop Hot Topics Manage. Internet, Cloud, Enterprise Netw. Services (Hot-ICE)*, 2012, pp. 1–6.

[6] D. Gligoroski and K. Kralevska, "Expanded combinatorial designs as tool to model network slicing in 5G," *IEEE Access*, vol. 7, pp. 54879–54887, 2019.

[7] *Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) Radio Access Capabilities, V 16.3.0*, 3GPP Standard TS 36.306, 2020.

[8] X. Zhang, X. Guo, and H. Zhang, "RB allocation scheme for eMBB and URLLC coexistence in 5G and beyond," *Wireless Commun. Mobile Comput.*, vol. 2021, Oct. 2021, Art. no. 6644323.

[9] M. Alsenwi, N. H. Tran, M. Bennis, S. R. Pandey, A. K. Bairagi, and C. S. Hong, "Intelligent resource slicing for eMBB and URLLC coexistence in 5G and beyond: A deep reinforcement learning based approach," *IEEE Trans. Wireless Commun.*, vol. 20, no. 7, pp. 4585–4600, Jul. 2021.

[10] A. K. Bairagi, M. S. Munir, M. Alsenwi, N. H. Tran, and C. S. Hong, "A matching based coexistence mechanism between eMBB and uRLLC in 5G wireless networks," in *Proc. 34th ACM/SIGAPP Symp. Appl. Comput.*, 2019, pp. 2377–2384.

[11] A. K. Bairagi et al., "Coexistence mechanism between eMBB and uRLLC in 5G wireless networks," *IEEE Trans. Commun.*, vol. 69, no. 3, pp. 1736–1749, Mar. 2021.

[12] A. Anand, G. De Veciana, and S. Shakkottai, "Joint scheduling of URLLC and eMBB traffic in 5G wireless networks," in *Proc. IEEE Conf. Comput. Commun.*, 2018, pp. 1970–1978.

[13] J. Tang, B. Shim, and T. Q. S. Quek, "Service multiplexing and revenue Maximization in sliced C-RAN incorporated with URLLC and multicast eMBB," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 4, pp. 881–895, Apr. 2019.

[14] R. Kassab, O. Simeone, and P. Popovski, "Coexistence of URLLC and eMBB services in the C-RAN uplink: An information-theoretic study," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, 2018, pp. 1–6.

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

[16] L. Marijanović, S. Schwarz, and M. Rupp, "Optimal resource allocation with flexible numerology," in *Proc. IEEE Int. Conf. Commun. Syst. (ICCS)*, 2018, pp. 136–141.

[17] L. Marijanovic, S. Schwarz, and M. Rupp, "A novel optimization method for resource allocation based on mixed numerology," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2019, pp. 1–6.

[18] T. T. Nguyen, V. N. Ha, and L. B. Le, "Wireless scheduling for heterogeneous services with mixed numerology in 5G wireless networks," *IEEE Commun. Lett.*, vol. 24, no. 2, pp. 410–413, Feb. 2020.

[19] "5G; study on new radio (NR) access technology," 3GPP, Sophia Antipolis, France, Rep. 38.912, 2018.

[20] L. Marijanović, S. Schwarz, and M. Rupp, "Multiplexing services in 5G and beyond: Optimal resource allocation based on mixed numerology and mini-slots," *IEEE Access*, vol. 8, pp. 209537–209555, 2020.

[21] Y. Prathyusha and T.-L. Sheu, "Coordinated resource allocations for eMBB and URLLC in 5G communication networks," *IEEE Trans. Veh. Technol.*, vol. 71, no. 8, pp. 8717–8728, Aug. 2022.

[22] M. Zambianco and G. Verticale, "Mixed-numerology interference-aware spectrum allocation for eMBB and URLLC network slices," in *Proc. 19th Mediterr. Commun. Comput. Netw. Conf. (MedComNet)*, 2021, pp. 1–8.

[23] M. Zambianco and G. Verticale, "A reinforcement learning agent for mixed-numerology interference-aware slice spectrum allocation with non-deterministic and deterministic traffic," *Comput. Commun.*, vol. 189, pp. 100–109, May 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0140366422000858

[24] M. Setayesh, S. Bahrami, and V. W. Wong, "Resource slicing for eMBB and URLLC services in radio access network using hierarchical deep learning," *IEEE Trans. Wireless Commun.*, vol. 21, no. 11, pp. 8950–8966, Nov. 2022.

[25] M. Mhedhbi, M. Morcos, A. Galindo-Serrano, and S. E. Elayoubi, "Performance evaluation of 5G radio configurations for industry 4.0," in *Proc. Int. Conf. Wireless Mobile Comput., Netw. Commun. (WiMob)*, 2019, pp. 1–6.

[26] D. Kotagiri, A. Sawabe, E. Takahashi, T. Iwai, T. Onishi, and Y. Nishikawa, "Context-based mixed-numerology profile selection for 5G and beyond," in *Proc. IEEE 19th Annu. Consum. Commun. Netw. Conf. (CCNC)*, 2022, pp. 611–616.

[27] K. Boutiba, M. Bagaa, and A. Ksentini, "Radio resource management in multi-numerology 5G new radio featuring network slicing," in *Proc. IEEE Int. Conf. Commun.*, 2022, pp. 359–364.

[28] A. Karimi, K. I. Pedersen, N. H. Mahmood, G. Pocovi, and P. Mogensen, "Efficient low complexity packet scheduling algorithm for mixed URLLC and eMBB traffic in 5G," in *Proc. IEEE 89th Veh. Technol. Conf. (VTC-Spring)*, 2019, pp. 1–6.

[29] A. Destounis and G. S. Paschos, "Complexity of URLLC scheduling and efficient approximation schemes," in *Proc. Int. Symp. Model. Optim. Mobile, Ad Hoc, Wireless Netw. (WiOPT)*, 2019, pp. 1–8.

[30] Y. Zhao, X. Chi, L. Qian, Y. Zhu, and F. Hou, "Resource allocation and slicing puncture in cellular networks with eMBB and URLLC terminals co-existence," *IEEE Internet Things J.*, vol. 9, no. 19, pp. 18431–18444, Oct. 2022.

[31] L.-H. Shen, C.-Y. Su, and K.-T. Feng, "CoMP enhanced subcarrier and power allocation for multi-numerology based 5G-NR networks," *IEEE Trans. Veh. Technol.*, vol. 71, no. 5, pp. 5460–5476, May 2022.

[32] M. Almekhlafi, M. A. Arfaoui, C. Assi, and A. Ghrayeb, "Superposition-based URLLC traffic scheduling in 5G and beyond wireless networks," *IEEE Trans. Commun.*, vol. 70, no. 9, pp. 6295–6309, Sep. 2022.

[33] W. Ning, Y. Wang, M. Liu, Y. Chen, and X. Wang, "Mission-critical resource allocation with puncturing in industrial wireless networks under mixed services," *IEEE Access*, vol. 9, pp. 21870–21880, 2021.

[34] M. Almekhlafi, M. A. Arfaoui, M. Elhattab, C. Assi, and A. Ghrayeb, "Joint resource allocation and phase shift optimization for RIS-aided eMBB/URLLC traffic multiplexing," *IEEE Trans. Commun.*, vol. 70, no. 2, pp. 1304–1319, Feb. 2022.

[35] G. Interdonato, S. Buzzi, C. D'Andrea, L. Venturino, C. D'Elia, and P. Vendittelli, "On the coexistence of eMBB and URLLC in multi-cell massive MIMO," 2023, *arXiv:2301.03575*.

[36] A. Esmaeily, K. Kralevska, and T. Mahmoodi, "Slicing scheduling for supporting critical traffic in beyond 5G," in *Proc. IEEE 19th Annu. Consum. Commun. Netw. Conf. (CCNC)*, 2022, pp. 637–643.

[37] *User Equipment (UE) Radio Transmission and Reception; Part 1: Range 1 Standalone*, 3GPP TS 36.101-1, R16, V16.0.0, Sep. 2020.

[38] "User equipment (UE) radio transmission and reception; part 4: Performance requirements," 3GPP, Sophia Antipolis, France, Rep. TR 38.101-4, V16.0.0, Mar. 2020.

[39] X. Lin, D. Yu, and H. Wiemann, "A primer on bandwidth parts in 5G new radio," 2020, *arXiv:2004.00761*.

[40] H. V. K. Mendis, P. E. Heegaard, V. Casares-Giner, F. Y. Li, and K. Kralevska, "Transient performance modelling of 5G slicing with mixed numerologies for smart grid traffic," in *Proc. IEEE 26th Int. Workshop Comput. Aided Model. Design Commun. Links Netw. (CAMAD)*, 2021, pp. 1–7.

[41] A. B. Kihero, M. S. J. Solaija, and H. Arslan, "Inter-numerology interference for beyond 5G," *IEEE Access*, vol. 7, pp. 146512–146523, 2019.

[42] X. Zhang, L. Zhang, P. Xiao, D. Ma, J. Wei, and Y. Xin, "Mixed numerologies interference analysis and inter-numerology interference cancellation for windowed OFDM systems," *IEEE Trans. Veh. Technol.*, vol. 67, no. 8, pp. 7047–7061, Aug. 2018.

[43] A. F. Demir and H. Arslan, "Inter-numerology interference management with adaptive guards: A cross-layer approach," *IEEE Access*, vol. 8, pp. 30378–30386, 2020.

[44] *5G NR; Base Station (BS) Radio Transmission and Reception*, 3GPP Standard TS 138 104, V15.2.0, Release 15, 2018.

[45] *Technical Specification Group Services and System Aspects; Service Requirements for Cyber-Physical Control Applications in Vertical Domains*, 3GPP Standard TS 22.104, V16.5.0, Sep. 2020.

[46] "Technical specification group services and system aspects," 3GPP, Sophia Antipolis, France, Rep. TR 21.915, V1.1.0, Release 15, Mar. 2019.

[47] K. Ying, J. M. Kowalski, T. Nogami, Z. Yin, and J. Sheng, "Coexistence of enhanced mobile broadband communications and ultra-reliable low-latency communications in mobile front-haul," in *Proc. Broadband Access Commun. Technol. XII*, Jan. 2018, Art. no. 105590C.

[48] M. Almekhlafi, M. A. Arfaoui, C. Assi, and A. Ghrayeb, "Joint resource and power allocation for URLLC-eMBB traffics multiplexing in 6G wireless networks," in *Proc. IEEE Int. Conf. Commun.*, 2021, pp. 1–6.

[49] T. L. Marzetta, *Fundamentals of Massive MIMO*. Cambridge, U.K.: Cambridge Univ. Press, 2016.

[50] R. Nissel and M. Rupp, "OFDM and FBMC-OQAM in doubly-selective channels: Calculating the bit error probability," *IEEE Commun. Lett.*, vol. 21, no. 6, pp. 1297–1300, Jun. 2017.

[51] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[52] J. Scarlett, V. Y. F. Tan, and G. Durisi, "The dispersion of nearest-neighbor decoding for additive non-Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 63, no. 1, pp. 81–92, Jan. 2017.

[53] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[54] M. Grant and S. Boyd. "CVX: MATLAB software for disciplined convex programming." Accessed: Dec. 2022. [Online]. Available: http://cvxr.com/cvx/

[55] U. Feige, M. Feldman, and I. Talgam-Cohen, "Oblivious rounding and the integrality gap," in *Proc. Approx. Randomization, Combinatorial Optim. Algorithms Techn. (APPROX/RANDOM)*, 2016, pp. 1–28.

[56] O. Ibe, *Markov Processes for Stochastic Modeling*. Oxford, U.K.: Newnes, 2013.

[57] K. S. Kim et al., "Ultrareliable and low-latency communication techniques for tactile Internet services," *Proc. IEEE*, vol. 107, no. 2, pp. 376–393, Feb. 2019.

[58] "Massive MIMO systems for 5G and beyond networks." Accessed: Dec. 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7284607/#:~:text=Theoretically%2C%20Massive%20MIMO%20systems%20can, in%20massive%20MIMO%20base%20station

[59] T. Bhattacharjee and M. Jamil, "GOOSE publishing and receiving operations of IEC 61850 enabled IEDs," in *Proc. IEEE 1st Int. Conf. Energy, Syst. Inf. Process. (ICESIP)*, 2019, pp. 1–6.

[60] "New services and applications with 5G ultra-reliable low latency communications." 5Gamericas. [Online]. Available: https://www.5gamericas.org/wp-content/uploads/2019/07/5G_Americas_URLLLC_White_Paper_Final__updateJW.pdf

[61] O. Aydin et al., "D4.1 draft air interface harmonization and user plane design," EU Project, Brussels, Belgium, document METIS-II/D4.1, May 2016.

[62] M. Alsenwi, N. H. Tran, M. Bennis, A. K. Bairagi, and C. S. Hong, "eMBB-URLLC resource slicing: A risk-sensitive approach," *IEEE Commun. Lett.*, vol. 23, no. 4, pp. 740–743, Apr. 2019.

[63] R. K. Jain, D.-M. W. Chiu, and W. R. Hawe, "A quantitative measure of fairness and discrimination," Eastern Res. Lab., Digit. Equip. Corp., Hudson, MA, USA, Rep. DEC-TR-301, 1984.

[64] H. Yin, L. Zhang, and S. Roy, "Multiplexing URLLC traffic within eMBB services in 5G NR: Fair scheduling," *IEEE Trans. Commun.*, vol. 69, no. 2, pp. 1080–1093, Feb. 2021.

[65] A. A. Esswie and K. I. Pedersen, "Multi-user preemptive scheduling for critical low latency communications in 5G networks," in *Proc. IEEE Symp. Comput. Commun. (ISCC)*, 2018, pp. 136–141.

ALI ESMAEILY (Student Member, IEEE) received the master's degree in telecommunications engineering from the Polytechnic University of Catalonia in 2018. He is currently pursuing the Ph.D. degree with the Department of Information Security and Communication Technology, Norwegian University of Science and Technology. His research interests include cloud computing, service orchestration, 5G NR, and network slicing.

H. V. KALPANIE MENDIS received the bachelor's degree (Hons.) in electrical and information engineering from the University of Ruhuna, Sri Lanka, in 2015, and the master's degree in information and communication technology from the University of Agder, Norway, with a focus on 5G ultra-reliable communications. She is currently pursuing the Ph.D. degree with the Norwegian University of Science and Technology. Her research interests lie in the areas of 5G end-to-end network slicing, intent-based networking, multi-RAT architectures, software-defined networking, network function virtualization, management and orchestration of networks, and dependability.

TOKTAM MAHMOODI (Senior Member, IEEE) received the B.Sc. degree in electrical engineering from the Sharif University of Technology, Iran, in 2002, and the Ph.D. degree in telecommunications from King's College London, U.K., in 2009, where she is currently the Head of the Centre for Telecommunications Research with the Department of Engineering. Her research interests include mobile communications, network intelligence, and low-latency networking.

KATINA KRALEVSKA (Member, IEEE) received the M.Sc. degree in mobile and wireless communications from Ss. Cyril and Methodius University, Skopje, Macedonia, in 2012, and the Ph.D. degree in telematics from the Norwegian University of Science and Technology, Trondheim, Norway, in 2016, where she has been an Associate Professor with the Department of Information Security and Communication Technology since 2018. She was the Deputy Head of the Department for two years from 2019 to 2020. Her research interests and activities lie in the areas of next-generation networks, coding theory, and blockchain.