Received 29 July 2020; revised 10 September 2020; accepted 30 September 2020. Date of publication 12 October 2020; date of current version 22 October 2020.

Digital Object Identifier 10.1109/OJCOMS.2020.3029553

Joint Power and Resource Block Allocation for Mixed-Numerology-Based 5G Downlink Under Imperfect CSI

PRAVEEN KUMAR KORRAI (Student Member, IEEE), EVA LAGUNAS (Senior Member, IEEE), ASHOK BANDI (Student Member, IEEE), SHREE KRISHNA SHARMA (Senior Member, IEEE), AND SYMEON CHATZINOTAS (Senior Member, IEEE)

SnT, University of Luxembourg, 1855 Luxembourg City, Luxembourg

CORRESPONDING AUTHOR: P. K. KORRAI (e-mail: praveen.korrai@uni.lu)

This work was supported by Luxembourg National Research Fund (FNR) through the AFR Grant for the Ph.D. Project LACLOCCN AFR under Grant 12561031.

ABSTRACT Fifth-generation (5G) of wireless networks are expected to accommodate different services with contrasting quality of service (QoS) requirements within a common physical infrastructure in an efficient way. In this article, we address the radio access network (RAN) slicing problem and focus on the three 5G primary services, namely, enhanced mobile broadband (eMBB), ultra-reliable and low-latency communications (URLLC) and massive machine-type communications (mMTC). In particular, we formulate the joint allocation of power and resource blocks to the heterogeneous users in the downlink targeting the transmit power minimization and by considering mixed numerology-based frame structures. Most importantly, the proposed scheme does not only consider the heterogeneous QoS requirements of each service, but also the queue status of each user during the scheduling of resource blocks. In addition, imperfect Channel State Information (CSI) is considered by including an outage probabilistic constraint into the formulation. The resulting non-convex problem is converted to a more tractable problem by exploiting Big-M formulation, probabilistic to non-probabilistic transformation, binary relaxation and successive convex approximation (SCA). The proposed solution is evaluated for different mixed-numerology resource grids within the context of strict slice-isolation and slice-aware radio resource management schemes via extensive numerical simulations.

INDEX TERMS RAN resource slicing, power minimization, resource block allocation, mixed-numerologies, eMBB, URLLC and mMTC.

I. INTRODUCTION

THE UPCOMING fifth-generation (5G) of wireless networks are expected to manage a wide variety of verticals with different requirements such as high data-rate, low-latency, and reliability [1]. According to International Telecommunication Union Recommendations (ITU-R), three generic usage scenarios, namely enhanced mobile broadband (eMBB), ultra-reliable low-latency communications (URLLC), and massive machine-type communications (mMTC) are expected in the 5G multi-service wireless network [2]. Out of these, eMBB is a direct extension of the current 4G broadband services, which requires stable connections with higher data rates to support applications like virtual reality (VR) [3] and high definition (HD) video [4]. URLLC aims to support the low latency transmissions of small packets with high reliability, and it covers applications such as industrial automation [5], vehicular communications [6], and real-time tactile Internet services [7]. Furthermore, mMTC [8] supports the services that connect a massive number of devices, where each device is intermittently active and sends small packets of data. It covers applications like smart cities and sensor networks for

farming [9]. However, the current one-size-fits-all network architecture is not sufficiently flexible to accommodate these services [10].

Accommodating eMBB, URLLC, and mMTC within the same radio access network (RAN) architecture while ensuring their potential co-existence is extremely challenging due to their contrasting quality-of-service (QoS) requirements. In 5G, to overcome this problem, different services are allowed to co-exist within the same network architecture by using the so-called network slicing mechanism [11]. Network slicing is a resource allocation mechanism that dynamically shares the available computing, communication, and storage resources among the existing services while guaranteeing their isolation and required performance levels. This slicing process can be executed on both the core network (CN) and the radio access network (RAN) to create the end-to-end logical networks or slices. In this article, we focus on the optimization of the RAN slicing mechanism, which involves the efficient allocation of physical radio resources available at the base station. In particular, we concentrate on the transmit power and the time-frequency unit or resource block (RB) assignment.

On the other hand, the conventional uniform RB grid definition considering single numerology currently utilized by the 4G Long-Term evolution (LTE) has been shown to be unsuitable for the heterogeneous verticals expected in 5G [12]. Since a standard underlying physical layer should provide a multitude of different services simultaneously, a separate radio architecture design for each service is not practical due to the complexity and additional cost. In addition, it is not viable to design a one-fits-all solution to achieve the requirements of all services [13]. To this end, flexible mixed numerologies have been recently proposed for the 5G air interface in 3GPP Release 15 [14]. As a result, each service can choose an appropriate numerology to allocate its data transmission while satisfying the stringent requirements of each service [15].

Although mixed-numerologies enable enhanced flexibility, they also introduce new challenges related to the RAN slicing mechanism. In particular, the dynamic allocation of mixed-numerology-based RB and associated transmit power in accordance with the instantaneous user traffic loads represents a major challenge. Furthermore, the effectiveness of different resource slicing strategies such as slice-isolation and slice-aware [16] need to be studied in the context of resource allocation for various services with varying traffic demands. Hence, in this article, inspired by [16], we firstly consider slice-isolation and slice-aware radio resource allocation schemes for allocating the RAN resources to users with heterogeneous requirements. Later, we address the associated resource allocation problems by jointly assigning the transmit power and radio RBs with mixed numerologies to the different users according to their requirements and queue status. In the following sub-sections, we briefly review the related works from the literature and highlight this article's contributions.

A. RELATED WORKS

Dynamic multiplexing of eMBB, URLLC, and mMTC services using the slicing based RAN resource allocation mechanisms has recently received a significant research interest in the literature. Some of the relevant studies are discussed briefly in the following.

The radio-channel and QoS aware resource scheduling technique has been proposed in [17] for the dynamic allocation of the radio resources to the eMBB and URLLC users according to their stringent QoS requirements. In [17], by interacting with the link adaptation process, the proposed scheduling mechanism adjusts the block error probability (BLEP) of URLLC transmissions by following the instantaneous traffic load in the cell. Also, a new channel quality (CQ) measuring mechanism has been investigated to enhance the efficiency of the URLLC link adaptation process. Furthermore, the authors of [18] have investigated the packet-size and control channel aware radio resource allocation mechanism to improve the performance of the scheduler proposed in [17]. A communication-theoretic model has been introduced in [19] to provide a view on non-orthogonal and orthogonal slicing of radio resources for the dynamic co-existence of eMBB, URLLC and mMTC services on the same physical network infrastructure. In [19], the authors have considered the puncturing technique to place the URLLC traffic on the ongoing eMBB transmissions. The joint scheduling algorithm for the URLLC and eMBB services has been designed in [20], where the URLLC data traffic is scheduled on the eMBB resources to satisfy the stringent latency requirement of URLLC service while maximizing the utility of eMBB. In [21], the authors have explored a risk-sensitive approach for the efficient scheduling of radio resources to the URLLC and eMBB service. Particularly, they considered a chance constraint to ensure the URLLC reliability and conditional value at risk function to measure the risk for eMBB users. Moreover, the aforementioned works [17]–[21] prioritize the URLLC service and schedule the URLLC traffic arrivals on the assigned resources of eMBB service that leads to performance degradation of eMBB in terms of data rate and reliability at the higher URLLC traffic. Also, the proposed puncturing based mechanisms increase the decoding complexity at the receiver and control channel (CCH) overhead due to the overlapping of services. Further, most of the works considered fixed numerology based time-frequency resources for the scheduling process.

On the other hand, the radio resource allocation mechanisms with mixed-numerologies in time and frequency domains have received increasing research attention due to its flexibility in supporting different services within the same transmit band. Optimization methods for the allocation of radio resources with variable RB structures (i.e., also referred to as frame tailing approach) to the services with different QoS requirements have been studied in [22], [23]. The authors of [22] considered the resource optimization problem for capacity enhancement, and in [23], the authors

investigated a resource allocation problem to maximize the number of admitted users. However, the proposed works [22], [23] assume that the time-frequency resources mapping process is to be done at the pre-processing stage, and it is decoupled from service scheduling. Further, at the dynamic traffic demands, this frame tailing approach substantially increases the co-existence overhead by improving the number of boundaries between numerologies.

Moreover, the works [13], [24] have investigated the multiplexing of services in the frequency, time domains to analyze the compatibility and support for the co-existence of different services. In [24], the authors have confirmed the efficiency of multiplexing of different numerologies in the frequency domain through experimented field trails for the performance evaluations of OFDM based 5G waveforms.

A multi-numerology and shortened time-slot based resource allocation scheme has proposed in [25] to assign the resources for different priority services. In this work, the authors assumed that each time-slot chooses single numerology according to the requirements of service type. The selection of unique numerology in every time-slot can avoid the effect of inter-numerology interference (INI) and the occurred spectrum wastage due to the utilization of guard-bands. By considering the RAN slices with different numerologies, the authors have analyzed the trade-off between the flexibility and overhead related to their potential co-existence in [26]. The slicing problem has been considered in [27] to show how time-frequency resources with different numerologies can be allocated to serve the users in a 5G system. In [28], a resource optimization method has investigated the allocation of radio resources with mixed-numerologies to multiple users from latency sensitive and latency tolerant services. Furthermore, in our previous work [29], we have investigated the slice-aware RAN resource allocation mechanism to optimize the LTE standard radio resources (i.e., 0.5ms each transmission time interval (TTI), 1ms sub-frame and 10ms frame) between eMBB and URLLC services according to their stringent QoS requirements. Also, in [29], we considered the modulation coding scheme (MCS) based link adaptation process and different queue models for the allocation of resources to eMBB and URLLC users.

However, the works mentioned earlier have not targeted to design the resource slicing algorithms for energy-efficient communication in OFDMA systems. Further, in the literature, many works extensively studied energy-efficient resource allocation algorithms in 4G communication scenarios. Reference [30] investigated a resource optimization problem for energy-efficient communication in OFDMA systems with limited backhaul capacity and shown the trade-off between energy efficiency and network backhaul capacities. In [31], the authors have studied the energy-efficient resource optimization in OFDMA heterogeneous networks by considering the minimum data rate, rate fairness, and BSs' transmit power budget constraints. However, the works in [30], [31] are not directly applicable to the upcoming 5G wireless networks, which support multiple services with heterogeneous requirements. Besides, the energy-efficient resource allocation algorithms have been majorly assumed to allocate resource blocks with fixed numerology to the users requesting a single service. In other words, the performance efficiency of mixed numerologies within the context of RAN resource allocation strategies for OFDMA systems has not been investigated until today. Furthermore, none of the works mentioned earlier studied the joint optimization of transmit power and RBs of mixed-numerologies to the eMBB, URLLC, and mMTC users while guaranteeing their stringent QoS requirements. Moreover, most of the earlier mentioned works did not consider the link adaptation process to select true MCS.

B. MAJOR CONTRIBUTIONS

Motivated by these aforementioned observations, in this work, we propose strict slice-isolation and slice-aware RAN slicing strategies to ensure the efficient co-existence of eMBB, URLLC, and mMTC services in the OFDMA network. Furthermore, we jointly allocate the transmission power and RBs with mixed-numerologies to the users by considering the MCS based link adaptation process, different traffic arrival patterns (i.e., queue models), and imperfect channel state information (CSI). Importantly, we formulate these RAN slicing problems as optimization problems to minimize the consumed power of BS while satisfying the isolation constraints, outage constraint, and QoS requirements of users requesting different services. The main contributions of the study are summarized as follows:

- Firstly, we formulate the joint power and RB allocation problem to minimize the overall transmit power at the BS while satisfying the isolation constraints, minimum Signal-to-Noise Ratio (SNR) constraints, and latency related constraints. Each service's reliability requirements are ensured via adaptive MCS tailored to the corresponding Block Error Probability (BEP). Due to the presence of product terms in the objective functions and binary optimization variable used for the allocation of RBs, the formulated optimization problems are highly non-convex and of combinatorial nature, which are very difficult to solve in polynomial time. Therefore, a low-complexity sub-optimal approach is proposed to jointly address the power and RBs allocation problem using the successive convex approximation (SCA) algorithm.
- Secondly, mixed-numerology based time-frequency resource grid models are considered to allocate the radio resources to the different users according to their queue status and stringent QoS requirements. Further, we compare the effectiveness of the different mixed-numerology based resource grid models with the conventional fixed-numerology based resource grid via numerical simulations.
- Thirdly, strict slice-isolation and slice-aware radio resource allocation mechanisms are considered for multiplexing the users from URLLC, mMTC, and

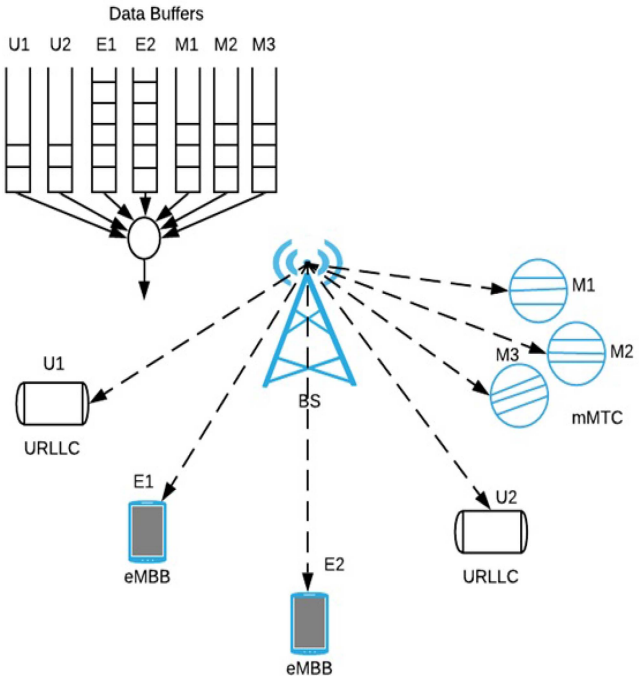


FIGURE 1. Illustration of DL single-cell cellular network serving URLLC, eMBB and mMTC users according to their data traffic storage buffers (or queue status).

eMBB services efficiently on the same RAN infrastructure. In the slice-isolation scheme, a dedicated number of resources are allocated to each slice, and no resource sharing option is available among them. On the contrary, in the slice-aware scheme, the slices share the available resources with the other slices by considering the slice-specific control parameters. Further, through the extensive simulation campaign, the performances of slice-isolated and slice-aware resource allocation mechanisms with different mixed-numerologies are compared in terms of achievable data rates, packet latencies, and queue status, respectively.

- Fourthly, to achieve robustness to partially known channel uncertainties, we model the imperfect CSI at the base station (BS) by including a probabilistic outage constraint into the problem formulation.

The remainder of the paper is organized as follows. The system model, channel model with and without CSI and time-frequency resource grid models are discussed in Section II. The power minimization problems based on slice-isolated and slice-aware radio resource allocation mechanisms are formulated in Section III. The joint solution to the power and RBs allocation is presented in Section IV. The numerical evaluations are provided in Section V and finally, conclusions are drawn in Section VI.

II. SYSTEM MODEL

We consider the orthogonal frequency division multiple access (OFDMA) downlink (DL) scenario, where a BS provides coverage to K number of eMBB users, L number of URLLC users, and M number of mMTC users (i.e., overall $K+L+M$ number of users) that are randomly located across the network area as shown in Fig. 1. The users associated with the eMBB service, indexed by $\mathcal{U}_E = \{1, 2, \ldots, K\}$ generate the continuous traffic (i.e., full-buffer) with infinite packet size. The users from URLLC service, indexed by $\mathcal{U}_L = \{1, 2, \ldots, L\}$ are assumed to generate bursts of small packets of B bytes according to the FTP3 model [2] with the arrival rate of λ_U [packets/sec]. The users belonging to mMTC service, indexed by $\mathcal{U}_M = \{1, 2, \ldots, M\}$ are assumed to generate bursts of small packets in the range of 20 bytes to 200 bytes by following the Pareto distribution [32]. The complete details of the traffic models are given in Section V. Also, we consider that all the data from higher layers are

TABLE 1. Summary of notations.

Symbol	Definition
$\mathcal{U}_E$	Set of available eMBB users
K	Total number of eMBB users in the network
$\mathcal{U}_L$	Set of available URLLC users
L	Total number of URLLC users in the network
$\mathcal{U}_M$	Set of available mMTC users
M	Total number of mMTC users in the network
λ_U	URLLC packets arrival rate (packets per ms)
B	URLLC packet size (in bytes)
T	Time-slot duration in ms
$\mathcal{F}_i$	No. of available RBs per TTI in a slice with numerology 'i'
$\mathcal{T}_i$	No. of available TTIs in a slice with numerology 'i'
$\mathcal{P}_{RB}$	No. of available REs per RB (t_i, f_i)
$\mathcal{R}^u_{t_i,f_i}$	Data bits of u^{th} user that can be sent on the RB (t_i, f_i)
$\gamma^u_{t_i,f_i}$	Received SNR of u^{th} user on the RB with imperfect CSI
$\mathcal{D}^u_{t_i,f_i}$	Achievable data rate of u^{th} user on RB with imperfect CSI
$\varphi^u_{t_i,f_i}$	Received SNR of u^{th} user on the RB with perfect CSI
$\mathcal{C}^u_{t_i,f_i}$	Scheduled data rate of u^{th} user on RB with perfect CSI
$x^u_{t_i,f_i}$	Binary decision variable
P_{max}	Available total power at BS
$p^u_{t_i,f_i}$	Allocated power to the RB (t_i, f_i)
$h^u_{t_i,f_i}$	Perfect CSI coefficient between the BS and u^{th} user on the RB (t_i, f_i)
$\hat{h}^u_{t_i,f_i}$	Imperfect CSI coefficient between the BS and u^{th} user on the RB (t_i, f_i)
$e^u_{t_i,f_i}$	Estimated CSI error between the BS and u^{th} user on the RB
ν	Path loss exponent
$d_{BS,u}$	distance between the BS and user
σ^2_o	Noise Power
σ^2_e	CSI error variance
$\mathcal{I}(\cdot)$	SE of the selected MCS
w_u	Number of available packets in u^{th} URLLC user's queue
τ_u	Number of available packets in u^{th} mMTC user's queue
Φ_l	Number of available RBs in the URLLC slice
Φ_m	Number of available RBs in the mMTC slice
γ^s_{th}	SNR threshold for service 's'
e_u	Required no. of RBs for u^{th} eMBB user
β_u	Required no. of RBs for u^{th} URLLC user
α_u	Required no. of RBs for u^{th} mMTC user
$\mathcal{B}_E$	BEP target for eMBB users
$\mathcal{B}_L$	BEP target for URLLC users
$\mathcal{B}_M$	BEP target for mMTC users
λ_1	Penalty parameter 1
λ_2	Penalty parameter 2
$Y(\cdot)$	Penalty function
$\nabla Y(\cdot)$	Gradient of $Y(\cdot)$
$Pr(\cdot)$	probability of $(\cdot)$
Δ_{out}	Channel outage probability
$F^{-1}(\cdot)$	Inverse CDF of $(\cdot)$

received at the BS and stored in their respective user-specific transmission buffers until they get served as shown in Fig. 1. Note that in this work we assume the optimal data buffer sizes that equal the bandwidth delay product and do not cause the buffer bloat problems.

The BS serves the set of all active users $\{1, 2, 3, \ldots, K + L + M\}$ in the cell by allocating the time-frequency radio-resource blocks (RBs) with optimized powers. The complete details of time-frequency frame numerologies are discussed in the following section.

A. TIME-FREQUENCY FRAME NUMEROLOGIES

Unlike the 4G LTE system, the 5G wireless system adopts the scalable numerologies to satisfy the QoS requirements imposed by different services include eMBB, URLLC, and mMTC. Due to the utilization of scalable numerologies, the 5G's RB can have the bandwidth equal to 2^{μ} times of RB's bandwidth in 4G LTE system (i.e., 180 KHz) and the time-slot interval equal to $T = \frac{1}{2^{\mu}}$ ms, where $\mu = \{0, 1, 2\}$ represents the numerology index. Also, each time-slot has 14 OFDM symbols in the time-duration of T ms. Further, the RB is defined as the minimum resource allocation unit that comprises of one mini-slot and 12 sub-carriers with SCS of $2^{\mu} \times 15$ kHz. Therefore, the bandwidth of an RB varies from 180 KHz to 720 KHz depending the value of μ (i.e., numerology index). The complete details of the scalable numerologies are also summarized in Table 2.

In addition, to satisfy the requirements of critical latency services, we consider a mini-slot based frame structure where each time-slot is further divided into two mini-slots. Each mini-slot consists of 7 OFDM symbols in the time interval of $T/2$ ms. The slot and mini-slot based frame structures are shown in Fig. 2.

Herein, we assume that eMBB, URLLC, and mMTC services are simultaneously provided by a single cell. According to [33], lower numerologies are more appropriate for mMTC due to its support for a large number of simultaneously connected devices with lower power. Intermediate numerologies are suitable for eMBB, which requires higher data rates. Higher numerologies are more suitable for latency-critical applications of the URLLC service since they transmit short bursts of data packets. Therefore, for the mixed-numerologies case, eMBB service would prioritize Numerology 2 with SCS of 30 kHz and 0.25ms of TTI duration (i.e., SCS index $\mu = 1$), mMTC service would prioritize Numerology 1 with SCS of 15 kHz and 0.5ms of TTI duration (i.e., SCS index $\mu = 0$), and URLLC service would prioritize Numerology 3 with SCS of 60 kHz and 0.125ms of TTI duration (i.e., SCS index $\mu = 2$).

1) MULTIPLEXING OF MIXED NUMEROLOGIES IN FREQUENCY DOMAIN

In this model, the available carrier bandwidth for the DL transmissions is divided into many bandwidth parts (BWPs). It depends on the requirements of assigned service to the BWP that selects the suitable numerology. Note that the

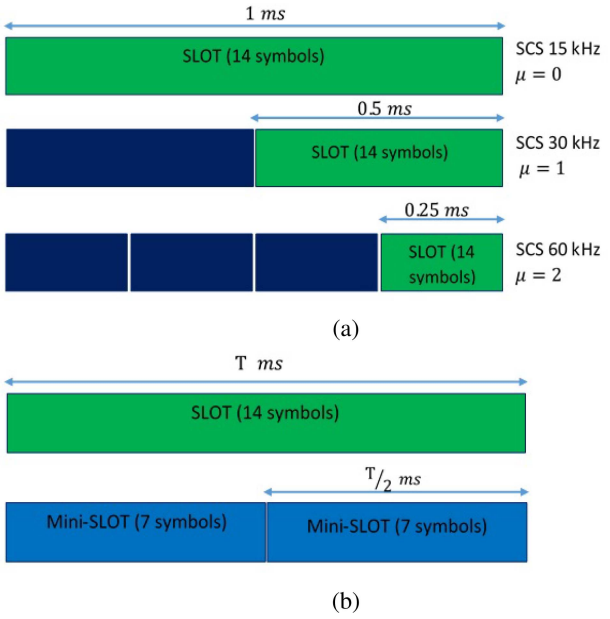


FIGURE 2. (a) Slot based frame and (b) Mini-slot based frame.

TABLE 2. 5G NR frame numerology.

Sub carrier spacing (SCS) Index (μ)	$\mu = 0$	$\mu = 1$	$\mu = 2$
Each SCS (in KHz)	15	30	60
Each RB's BW (in KHz)	180	360	720
Each time-slot's duration (in ms)	1	0.5	0.25
Time-duration of mini-slot (in ms)	0.5	0.25	0.125

BWP represents a dedicated time-frequency resource slice for a specific service. In this article, we assumed that proper BWP design have been already established based on the expected usage of the BS.

According to the chosen numerology, each BWP is partitioned into $\mathcal{F}_i$ number of sub-bands, indexed by $f_i = \{1, 2, \ldots, \mathcal{F}_i\}$ in the frequency-domain and $\mathcal{T}_i$ number of TTIs, indexed by $N_i = \{1, 2, \ldots, \mathcal{T}_i\}$ in the time-domain. Therefore, a total $\mathcal{F}_i \times \mathcal{T}_i$ number of RBs are available for the service using the i^{th} numerology. While this scheme does not require strict time synchronization procedures, the INI arises due to the utilization of different numerologies in the adjacent sub-bands. Thus, a fixed guard band is placed between the two adjacent numerologies (i.e., sub-bands) to suppress INI. This time-frequency resource grid model is shown in Fig. 3(a).

2) MULTIPLEXING OF MIXED NUMEROLOGIES IN TIME-DOMAIN

In this model, the multiplexing of the different services is done in time-domain, where every sub-frame of a frame is assigned to a specific service. Further, each sub-frame selects unique numerology to achieve the requirements of assigned service as shown Fig. 3(b). Similar to the frequency-domain multiplexing, we assume that the time division of the frame into different numerologies is already optimized in advanced based on the expected usage of the BS.

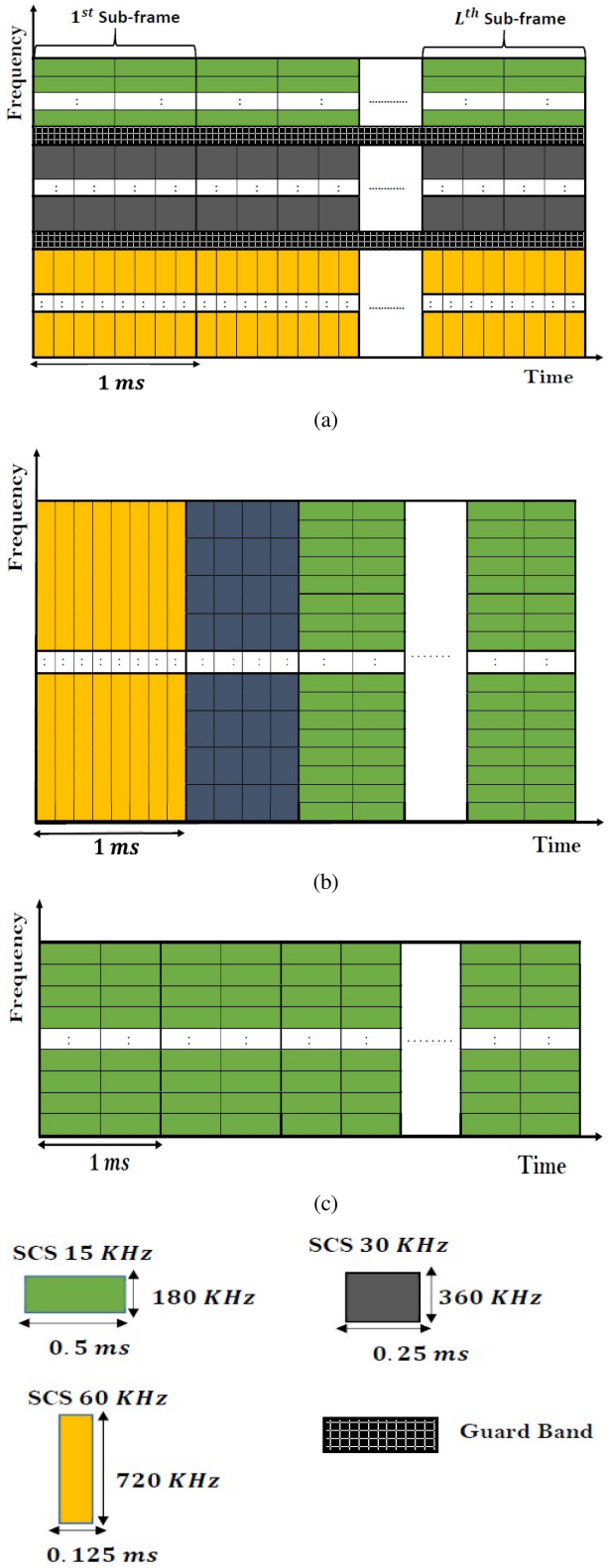


FIGURE 3. (a) Different numerologies in Frequency-domain, (b) different numerologies in time-domain and (c) Fixed Numerology.

This model suppresses the INI and wastage of spectrum due to the implementation of guard bands in the adjacent numerologies. However, the multiplexing in time translates into some time periods between numerologies which may negatively affect the performance of certain services (e.g., URLLC latency requirements may not be satisfied if $\mu = 2$-numerology is widely separated in time.

Now by using the frame numerology, the available transmission bandwidth is divided into multiple sub-bands, and the time-duration is slotted into multiple mini-slots (or TTIs), as shown in Fig. 3(b). Further, the same frame numerology assignment process is applied to all the sub-frames of a frame according to the assigned service. In this model, the utilization of the same RB structure for the complete duration of the sub-frame maintains the orthogonality between the consecutive RBs.

By following the chosen numerology, the available time-frequency resources in the sub-frame are divided into $\mathcal{F}_i \times \mathcal{T}_i$ number of RBs. The sets of sub-bands and time-slots are indexed by $f_i = \{1, 2, 3, \ldots, \mathcal{F}_i\}$ and $N_i = \{1, 2, 3, \ldots, \mathcal{T}_i\}$. Note that each sub-frame represents a dedicated resource slice for a particular service.

3) FIXED NUMEROLOGY

Fixed numerology based resource grid is currently used in the 4G LTE systems, where each RB comprises of 12 consecutive sub-carriers (i.e., for a complete bandwidth of 180KHz with 15KHz SCS) and 7 OFDM symbols in a time-slot of 0.5 ms. It is also one of the candidate numerologies for the upcoming 5G NR systems (i.e., $\mu = 0$).

In this model, the available frequency bandwidth for the DL transmissions is divided into F sub-bands indexed by $f = \{1, 2, \ldots, F\}$ and the time interval is slotted into transmission time intervals (TTIs) indexed by $t = \{1, 2, \ldots, N\}$ as shown in Fig. 3. The complete resource time-frequency resource grid consists of $F \times N$ numbers RBs, where each RB from the resource grid utilizes the same numerology with a fixed SCS and a fixed time-slot duration as shown in Fig. 3 (c).

B. CHANNEL MODEL

In most of the previous works, the resource scheduling mechanism is investigated by assuming the perfect CSI at the BS. However, in practice, due to some limitations such as limited (or delayed) feedback, quantization errors, CSI estimation errors, it is challenging to obtain perfect CSI at the BS. Hence, in this work, we propose a new optimization framework which takes into account the imperfect CSI.

In particular, we assume that the imperfect small scale fading channel is estimated at the BS before performing the resources optimization. In the following, since the large scale fading path loss and shadowing parameters are slowly varying and we assume that these can be estimated at the BS perfectly [34]. By utilizing the minimum mean square channel estimation error technique [35], the perfect channel coefficient between the BS and scheduled user on the RB (t_i, f_i) of numerology i is modeled as

$$h^u_{t_i,f_i} = \hat{h}^u_{t_i,f_i} + e^u_{t_i,f_i} \tag{1}$$

TABLE 3. Modulation and coding schemes (MCS) for eMBB, URLLC and mMTC services with different BEPs.

MCS	Modulation	SNR Th. [dB] $\mathcal{B}_M = 0.1$	SNR Th. [dB] $\mathcal{B}_E = 0.001$	SNR Th. [dB] $\mathcal{B}_U = 0.00001$	SE [bits/Symbol]
MCS1	QPSK	-6.5	-2.5	1.5	0.15
MCS2	QPSK	-4.0	0.0	4.0	0.23
MCS3	QPSK	-2.6	1.4	5.4	0.38
MCS4	QPSK	-1.0	3.0	7.0	0.60
MCS5	QPSK	1.0	5.0	9.0	0.88
MCS6	QPSK	3.0	7.0	11.0	1.18
MCS7	16QAM	6.6	10.6	14.6	1.48
MCS8	16QAM	10.0	14.0	18.0	1.91
MCS9	16QAM	11.4	15.4	19.4	2.41
MCS10	64QAM	11.8	15.8	19.8	2.73
MCS11	64QAM	13.0	17.0	21.0	3.32
MCS12	64QAM	13.8	17.8	21.8	3.90
MCS13	64QAM	15.6	19.6	23.6	4.52
MCS14	64QAM	16.8	20.8	24.8	5.12
MCS15	64QAM	17.6	21.6	25.6	5.55

where $\hat{h}^u_{t_i,f_i} \sim \mathcal{CN}(0, 1-\sigma_e^2)$ and $e^u_{t_i,f_i} \sim \mathcal{CN}(0, \sigma_e^2)$ represent the estimated CSI and estimated error, respectively. When the BS acquires the perfect CSI, the achievable SNR of the user 'u' on the RB (t_i, f_i) is computed as

$$\varphi^u_{t_i,f_i} = \frac{\left|h^u_{t_i,f_i}\right|^2 p^u_{t_i,f_i} d^{-v}_{BS,u}}{\sigma_o^2} \tag{2}$$

where $p^u_{t_i,f_i}$ is the allocated power to the each RB, $d_{BS,u}$ is the distance between the user and the BS, v is the path loss exponent, and σ_o^2 is the noise power. Now, according to Shannon's theory, the maximum scheduled data rate of the user 'u' on the RB (t_i, f_i) is expressed as

$$\mathcal{C}^u_{t_i,f_i} = \mathcal{B}^u_{t_i,f_i} \cdot \log_2\left(1 + \varphi^u_{t_i,f_i}\right) \tag{3}$$

where $\mathcal{B}^u_{t_i,f_i}$ is the bandwidth of the RB with numerology 'i'. Further, the achievable SNR of the user 'u' on the RB (t_i, f_i) with imperfect CSI can be computed

$$\gamma^u_{t_i,f_i} = \frac{S^u_{t_i,f_i}}{\sigma_o^2} \tag{4}$$

where $S^u_{t_i,f_i} = |\hat{h}^u_{t_i,f_i}|^2 p^u_{t_i,f_i} d^{-v}_{BS,u}$. The achievable data rate of the user 'u' on the RB with the imperfect CSI is written as

$$\mathcal{D}^u_{t_i,f_i} = \mathcal{B}^u_{t_i,f_i} \cdot \log_2\left(1 + \gamma^u_{t_i,f_i}\right) \tag{5}$$

When the resource scheduling process is executed based on the imperfect estimated CSI, the scheduled data rate may exceed Shannon's data rate (i.e., maximum data rate). In this scenario, the transmission may fail, which results in data outage [36]. Therefore, the outage probability is considered as a metric to measure the performance of case when the achievable data rate with imperfect CSI exceeds the scheduled data rate with perfect CSI (i.e., Shannon's data rate). Further, the outage probability condition imposed by imperfect CSI can be expressed as

$$\Pr\left(\mathcal{C}^u_{t_i,f_i} < \mathcal{D}^u_{t_i,f_i}\right) \leq \Delta_{out} \tag{6}$$

where $0 < \Delta_{out} < 1$.

C. LINK ADAPTATION PROCESS

In this process, first each user estimates the channel quality indicators (CQIs) for all the available RBs and feeds back its CQIs to the BS. If the RB is scheduled to the particular user, the MCS method enables the wireless system to select the suitable coding scheme according to the obtained CQI. Based on the BEP, and the received CQI feedback from the user, the minimum SNR threshold is set to achieve the appropriate MCS. Now, the data bits that can be transmitted on each RB of numerology i is computed as [29]

$$\mathcal{R}^u_{t_i,f_i} = \mathcal{P}_{RB} \cdot \mathcal{I}\left(\gamma^u_{t_i,f_i}\right) \tag{7}$$

where $\mathcal{P}_{RB}$ is the number of resource elements (REs) per RB after accounting the reference signals, and $\mathcal{I}(.)$ is the spectral efficiency (SE) of the selected MCS from Table 3 according to the estimated SNR. The reliability requirements of eMBB, URLLC, and mMTC are different from each other. Therefore, in our study, three distinct MCSs are considered, and the corresponding SNR levels of MCS for different BEP targets are given in Table 3 [29], [37]. Particularly, BEPs $\mathcal{B}_M = 10^{-1}$, $\mathcal{B}_E = 10^{-3}$, and $\mathcal{B}_U = 10^{-5}$ are considered for mMTC, eMBB and URLLC, respectively.

Then, the bit-rate of user that belongs to the service eMBB or URLLC or mMTC can be computed as

$$\mathbf{R}^u_s = \sum_{i\in\{1,2,3\}} \sum_{t_i=1}^{\mathcal{T}_i} \sum_{f_i=1}^{\mathcal{F}_i} x^u_{t_i,f_i} \mathcal{R}^u_{t_i,f_i}, \quad \begin{matrix} \forall u \in \mathcal{U}_E : s = mbb \\ \forall u \in \mathcal{U}_L : s = llc \\ \forall u \in \mathcal{U}_M : s = mtc. \end{matrix} \tag{8}$$

III. POWER MINIMIZATION BASED RADIO RESOURCE SLICING MECHANISMS

In this section, the following two approaches are considered for slicing the RAN resources at a sub-frame level for services with different QoS requirements.

- *Strict slice-isolation:* In this approach, a dedicated number of radio resources are assigned to slices without sharing options among services. In other words, each service is associated to a specific numerology, and cannot operate in other numerologies.

- *Slice-aware:* In this approach, slices share the available radio resources with other slices under specific constraints. These constraints ensure the performance guarantees of individual slices without adversely affecting each other's performance. Therefore, the services would be preferably assigned to a specific numerology but their operation in other numerologies is allowed.

Although the complexity of implementation and management of the slice-aware method is higher as compared to the slice-isolation scheme, it improves the resource utilization by efficiently allocating them to services according to its dynamic traffic arrivals.

By invoking the aforementioned mechanisms, the available RAN radio resources in the network are divided into multiple resource slices where each resource slice is assigned to the specific service. Further, the available RAN radio resources of each slice are allocated to requested users according to its traffic arrivals. Notice that the RAN radio resources considered for allocation in this work are transmit power and RBs.

In this work, we are interested in investigating (1) how to efficiently schedule the resources among eMBB, URLLC and mMTC to minimize the consumed power at the BS, thereby maximizing the energy efficiency while ensuring their specific QoS requirements; (2) how much transmit power is needed at the BS to support the users requesting different services in the network to achieve their specifications under the consideration of CSI imperfections and different traffic loads. Motivated by these interests, the joint power and RBs allocation problems are formulated as power-minimization problems in this section.

Furthermore, the resource allocation is performed at the sub-frame level. Thus, each sub-frame has $\sum_i(\mathcal{F}_i \times \mathcal{T}_i)$ number of RBs for the selected time-frequency resource grid. It is important to note here that the RBs can have different frame numerologies (i.e., SCS 30 KHz) to satisfy the requirements of users.

In the power minimization problem, the joint allocation of RBs and power to the users is considered with the objective of minimizing the total transmit power consumption at the BS subject to the following constraints:

- *Scheduling constraint*: With the help of a Boolean variable $x^u_{t_i,f_i}$ which represents the association between the user 'u' and RB (t_i, f_i) of numerology i, the binary scheduling constraint is mathematically formulated as,

$$x^u_{t_i,f_i} = \left\{ \begin{array}{ll} 1; & \text{If RB } (t_i,f_i) \text{ is allocated to user '}u\text{'} \\ 0; & \text{Otherwise} \end{array} \right\} \tag{C1}$$

- *Orthogonality constraint:* To comply with the requirement of OFDMA systems that an RB can only be allocated to a single user, we impose the following orthogonality constraint as

$$\sum_{u=1}^{K+L+M} x^u_{t_i,f_i} \leq 1; \forall i, t_i, f_i, u. \tag{C2}$$

- *Slice-isolation constraint:* In strict slice-isolation based resource allocation scheme, sharing of resources among slices is not allowed; this is compactly referred to as slice isolation constraint. By using $x^u_{t_i,f_i}$, the slice isolation constraint is formulated as

$$x^u_{t_i,f_i} = 0; \left\{ \begin{array}{c} \forall u \in \{\mathcal{U}_E \cup \mathcal{U}_L\}, t_3, f_3 \text{ or} \\ \forall u \in \{\mathcal{U}_M \cup \mathcal{U}_L\}, t_2, f_2 \text{ or} \\ \forall u \in \{\mathcal{U}_E \cup \mathcal{U}_M\}, t_1, f_1. \end{array} \right\} \tag{C3}$$

- *Channel outage constraint*: To account for the channel imperfections, the channel outage probability constraint is considered per RB as given below,

$$\Pr\left(\mathcal{C}^u_{t_i,f_i} < \mathcal{D}^u_{t_i,f_i}\right) \leq \Delta_{out}, \forall u, t_i, f_i, i. \tag{C4}$$

where $\Pr(\mathcal{C}^u_{t_i,f_i} < \mathcal{D}^u_{t_i,f_i})$ represents the probability of the achievable data rate of the user on the scheduled RB with perfect CSI is lesser than the achievable data rate of the user on the scheduled RB with imperfect CSI.

- *Minimum SNR constraint per RB:* Every active user from a URLLC or eMBB or mMTC service has to at least achieve a corresponding minimum level of SNR (i.e., $\gamma^s_{th}, s \in \{E, L, M\}$) on the assigned RB; this is simply referred to as minimum SNR constraint per RB. The minimum SNR constraint per RB is expressed as,

$$\gamma^u_{t_i,f_i} \geq x^u_{t_i,f_i}\gamma^s_{th}; \left\{ \begin{array}{l} \forall u \in \mathcal{U}_M, t_3, f_3, \\ \forall u \in \mathcal{U}_L, t_1, f_1, \\ \forall u \in \mathcal{U}_E, t_2, f_2. \end{array} \right\} \tag{C5}$$

This constraint helps the user to transmit the required number of bits on the scheduled RB according to its requesting service. Note that this constraint can be modified later on in the text by taking the outage requirement into account.

- *QoS Constraint for eMBB:* Every scheduled eMBB user must receive at least e_u number of RBs for every sub-frame to satisfy its overall minimum data rate requirement; this is simply referred to as QoS Constraint for eMBB. The QoS Constraint for eMBB is expressed as

$$\sum_{t_2=1}^{\mathcal{T}_2}\sum_{f_2=1}^{\mathcal{F}_2} x^u_{t_2,f_2} \geq e_u; \forall u \in \mathcal{U}_E \tag{C6}$$

- *QoS Constraint for URLLC*: Every scheduled URLLC should receive at least β_u number of RBs from the dedicated URLLC slice to vacate the available packets in the queues of URLLC users; this is referred to as QoS constraint for URLLC, which is expressed as

$$\sum_{t_1=1}^{\mathcal{T}_1}\sum_{f_1=1}^{\mathcal{F}_1} x^u_{t_1,f_1} \geq \beta_u; \forall u \in \mathcal{U}_L \tag{C7}$$

where $\beta_u = \min(\Omega_u, w_u)$, w_u is the number of available packets in u^{th} URLLC user's queue and Ω_u is the utmost

available number of RBs for each URLLC user. Further, Ω_u is obtained by

$$\Omega_u = \frac{w_u}{\sum_{u=1}^{L} w_u} \Phi_l \tag{9}$$

where Φ_l represents the number of available RBs in the URLLC slice.

- *QoS Constraint for mMTC:* Every scheduled mMTC user should receive at least α_u number of RBs from the dedicated mMTC slice to transmit the available data packets of mMTC users; this is called as QoS constraint for mMTC, which is written as

$$\sum_{t_3=1}^{\mathcal{T}_3} \sum_{f_3=1}^{\mathcal{F}_3} x^u_{t_3,f_3} \geq \alpha_u; \forall u \in \mathcal{U}_M \tag{C8}$$

where $\alpha_u = \min(\psi_u, \tau_u)$, τ_u is the number of available packets in u^{th} mMTC user's queue and ψ_u is the at most available number of RBs for each mMTC user. Further, ψ_u is obtained by

$$\psi_u = \frac{\tau_u}{\sum_{u=1}^{M} \tau_u} \Phi_m \tag{10}$$

where Φ_m represents the number of available RBs in the mMTC slice.

- *Power-related constraints:* The transmit power of the scheduled RB should be positive, and the sum of assigned transmit powers to all the scheduled RBs should not be more than the maximum available transmission power (i.e., budget power) at the BS; these are referred to as power-related constraints, which are expressed as

$$p^u_{t_i,f_i} \geq 0; \forall t_i, f_i, u \tag{C9}$$

$$\sum_{i\in\{1,2,3\}} \sum_{u=1}^{K+L+M} \sum_{f_i=1}^{\mathcal{F}_i} x^u_{t_i,f_i} p^u_{t_i,f_i} \leq P_{Max}, \forall t_i \tag{C10}$$

The objective function, the overall power consumption of the BS for each sub-frame, is then given by

$$\mathcal{P}_{tot} = \sum_{i\in\{1,2,3\}} \sum_{t_i=1}^{\mathcal{T}_i} \sum_{f_i=1}^{\mathcal{F}_i} \sum_{u=1}^{K+L+M} x^u_{t_i,f_i} p^u_{t_i,f_i}. \tag{11}$$

A. SLICE-ISOLATION BASED POWER AND RB ALLOCATION PROBLEM

The complete slice-isolation based power minimization problem is formulated as

$$\textbf{(P1:)} \quad \min_{\left\{x^u_{t_i,f_i},\ p^u_{t_i,f_i}\right\}} \quad \mathcal{P}_{tot}$$

subject to

$$(C1): x^u_{t_i,f_i} \in \{0,1\}; \forall i, t_i, f_i, u$$

$$(C2): \sum_{u=1}^{K+L+M} x^u_{t_i,f_i} \leq 1; \forall i, t_i, f_i$$

$$(C3): x^u_{t_i,f_i} = 0; \begin{cases} \forall u \in \{\mathcal{U}_M \cup \mathcal{U}_L\}, t_2, f_2; \\ \forall u \in \{\mathcal{U}_E \cup \mathcal{U}_L\}, t_3, f_3; \\ \forall u \in \{\mathcal{U}_E \cup \mathcal{U}_M\}, t_1, f_1; \end{cases}$$

$$(C4): \Pr\left(\mathcal{C}^u_{t_i,f_i} < \mathcal{D}^u_{t_i,f_i}\right) \leq \Delta_{out}, \forall u, t_i, f_i, i;$$

$$(C5): \gamma^u_{t_i,f_i} \geq x^u_{t_i,f_i} \gamma^s_{th}; \begin{cases} \forall u \in \mathcal{U}_M, t_3, f_3; \\ \forall u \in \mathcal{U}_L, t_1, f_1; \\ \forall u \in \mathcal{U}_E, t_2, f_2; \end{cases}$$

$$(C6): \sum_{t_2=1}^{\mathcal{T}_2} \sum_{f_2=1}^{\mathcal{F}_2} x^u_{t_2,f_2} \geq e_u; \forall u \in \mathcal{U}_E$$

$$(C7): \sum_{t_1=1}^{\mathcal{T}_1} \sum_{f_1=1}^{\mathcal{F}_1} x^u_{t_1,f_1} \geq \beta_u; \forall u \in \mathcal{U}_L$$

$$(C8): \sum_{t_3=1}^{\mathcal{T}_3} \sum_{f_3=1}^{\mathcal{F}_3} x^u_{t_3,f_3} \geq \alpha_u; \forall u \in \mathcal{U}_M$$

$$(C9): p^u_{t_i,f_i} \geq 0; \forall t_i, f_i, u$$

$$(C10): \sum_{i\in\{1,2,3\}} \sum_{u=1}^{K+L+M} \sum_{f_i=1}^{\mathcal{F}_i} x^u_{t_i,f_i} p^u_{t_i,f_i} \leq P_{max}, \forall t_i \tag{12}$$

Remarks:

- The value of SNR threshold γ^s_{th} requires to select from the MCS table according to the individual requirement of each service.
- The constraints (C5) and (C6) help the eMBB users to maintain the minimum rate requirement for every sub-frame. For instance, assume that the feasible SNR threshold (i.e., γ^E_{th} in (C4)) for eMBB service is 17.8dB, the minimum number of required RBs (i.e., e_u in (C5)) for each eMBB user is 5, MCS13 is the respective coding scheme to the provided SNR threshold, and the SE of MCS13 is 3.90 bits/symbol. As a result, every user from the eMBB service transmits at least $60 \times 3.90 = 234$ bits per RB. Finally, each user can transmit at least $234 \times 5 = 1170$ bits for a sub-frame. Similarly, the constraints (C5) and (C7) together take the advantage to vacate the queues of URLLC users, and also, the combination of constraints (C5) and (C9) satisfies the QoS requirement of mMTC users.
- We point out that the considered optimization problem might be infeasible because of the constraints related to the QoS requirements. In this case, an appropriate parameter setting for such constraints is essential to obtain a feasible problem formulation. We choose the parameters γ^s_{th}, e_u, β_u, and α_u appropriately as like [23]. Importantly, it is essential to consider admission control in addition to the power and RBs assignment into the problem formulation to avoid the issue of infeasibility, but this is out of the scope of this article.

The strict slice-isolation based radio resource allocation mechanism cannot fully utilize the the available radio resources of a slice under low traffic conditions. Moreover,

this mechanism does not share the underutilized resources with other slices that strive for more resources. Therefore, in the following section, we also formulate the slice-aware radio resource allocation mechanism to improve the utilization of radio-resources.

B. SLICE-AWARE POWER AND RB ALLOCATION PROBLEM

In this approach, as mentioned in the previous section, slices share the available radio resources with the other slices. Therefore, the constraint (C3) for the slice-isolation is not required for this optimization problem. Furthermore, when the scheduled URLLC user requires some additional RBs because of its traffic demand, then it takes the required number of RBs from other slices. In order to get some additional RBs from the other slices, the following constraint is considered for every user as,

$$\sum_{t_i=1}^{\mathcal{T}_i}\sum_{f_i=1}^{\mathcal{F}_i} x^u_{t_i,f_i} \geq \zeta_{i,u};\ \forall u \in \mathcal{U}_L, i \in \{2,3\} \qquad \text{(C11)}$$

where $\zeta_u = \min(\kappa, \xi_u)$ for the mixed numerologies in time domain, and $\zeta_u = \lfloor \min(\kappa, \xi_u)/2 \rfloor$ for the mixed numerologies in frequency domain, where $\xi_u = \max(0, w_u - \Omega_u)$ and κ is a threshold value that is provided in the simulation section.

Similarly, for the mMTC users, in order to access the RBs of other slices, the constraint (C12) is written as

$$\sum_{t_i=1}^{\mathcal{T}_i}\sum_{f_i=1}^{\mathcal{F}_i} x^u_{t_i,f_i} \geq \eta_u;\ \forall u \in \mathcal{U}_M, i \in \{1,2\} \qquad \text{(C12)}$$

where $\eta_u = \min(\rho, \chi_u)$ for the mixed numerologies in time domain, and $\eta_u = \lfloor \min(\rho, \chi_u)/2 \rfloor$ for the mixed numerologies in frequency domain, where $\chi_u = \max(0, \tau_u - \psi_u)$. In order to improve the data rates, the eMBB users may access the underutilized resources of other slices using the following constraint as

$$\sum_{t_i=1}^{\mathcal{T}_i}\sum_{f_i=1}^{\mathcal{F}_i} x^u_{t_i,f_i} \geq \mathcal{E}_{i,u};\ \forall u \in \mathcal{U}_E, i \in \{1,3\} \qquad \text{(C13)}$$

where $\mathcal{E}_{1,u} = \max(0, \lfloor \frac{\Phi_U - \sum_{u=1}^{L}\beta_u - \sum_{u=1}^{M}\eta_u}{K} \rfloor)$, and $\mathcal{E}_{3,u} = \max(0, \lfloor \frac{\Phi_M - \sum_{u=1}^{L}\zeta_u - \sum_{u=1}^{M}\alpha_u}{K} \rfloor)$.

The complete slice-aware resource allocation problem is mathematically formulated as

$$\begin{aligned}
(\mathbf{P2:}) \quad & \min_{\left\{x^u_{t_i,f_i},\ p^u_{t_i,f_i}\right\}} \quad \mathcal{P}_{tot} \\
& \text{subject to} \\
& (C1), (C2), (C4) - (C10) \text{ in } (P1) \\
& (C11): \sum_{t_i=1}^{\mathcal{T}_i}\sum_{f_i=1}^{\mathcal{F}_i} x^u_{t_i,f_i} \geq \zeta_u;\ \forall u \in \mathcal{U}_L, i \in \{2,3\} \\
& (C12): \sum_{t_i=1}^{\mathcal{T}_i}\sum_{f_i=1}^{\mathcal{F}_i} x^u_{t_i,f_i} \geq \eta_u;\ \forall u \in \mathcal{U}_M, i \in \{1,2\} \\
& (C13): \sum_{t_i=1}^{\mathcal{T}_i}\sum_{f_i=1}^{\mathcal{F}_i} x^u_{t_i,f_i} \geq \mathcal{E}_{i,u};\ \forall u \in \mathcal{U}_E, i \in \{1,3\}
\end{aligned} \qquad (13)$$

The formulated optimization problems (**P1**) and (**P2**) are combinatorial and highly non-convex due to the binary constraint $(C1)$ and the objective functions, probability constraint $(C4)$, and constraint $(C10)$, respectively. These problems are generally referred to as mixed-integer non-convex optimization problems, which are not easy to solve optimally in polynomial time. In general, the exhaustive search or brute-force method is essential for attaining the globally optimal solution (i.e., close to the analytical solution). However, such a technique has exponential complexity with respect to RBs, which is computationally infeasible even for lower dimensions. In order to avoid the higher computational complexities, in the next section, we propose low-complex sub-optimal solutions based on a successive convex approximation to the aforementioned problems.

IV. SOLUTION TO POWER MINIMIZATION BASED RESOURCE ALLOCATION SCHEMES

In this section, we provide solutions to the aforementioned slice-isolation and slice-aware resource optimization problems, where each solution comprises of four major steps as outlined in the following.

A. SOLUTION TO SLICE-ISOLATION PROBLEM

- *Step 1 (Big-M Formulation):* A reason for the non-convexity of (P1) is the presence of product term (i.e., $x^u_{t_i,f_i}p^u_{t_i,f_i}$) in objective function and constraint (C10). Towards addressing this issue, we introduce the following scheduling constraint based on big-M formulation theory to ensure that if $x^u_{t_i,f_i} = 0$, then $p^u_{t_i,f_i} = 0$ [38],

$$0 \leq p^u_{t_i,f_i} \leq x^u_{t_i,f_i} P_{max} \qquad (14)$$

By using (14), the optimization problem $(P1)$ is reformulated as

$$\begin{aligned}
(\mathbf{P1a:}) \quad & \min_{\left\{x^u_{t_i,f_i},\ p^u_{t_i,f_i}\right\}} \quad \hat{\mathcal{P}}_{tot} \\
& \text{subject to} \\
& (C1) - (C8) \text{ in } (P1) \\
& (C9): 0 \leq p^u_{t_i,f_i} \leq x^u_{t_i,f_i} P_{Max};\ \forall t_i, f_i, u \\
& (C10): \sum_{i\in\{1,2,3\}}\ \sum_{u=1}^{K+L+M}\ \sum_{f_i=1}^{\mathcal{F}_i} p^u_{t_i,f_i} \leq P_{max},\ \forall t_i
\end{aligned} \qquad (15)$$

where $\hat{\mathcal{P}}_{tot} = \sum_{i\in\{1,2,3\}}\sum_{t_i=1}^{\mathcal{T}_i}\sum_{f_i=1}^{\mathcal{F}_i}\sum_{u=1}^{K+L+M} p^u_{t_i,f_i}$.

Remarks:

– The constraint (C9) is valid for both $x^u_{t_i,f_i} = 1$ and $x^u_{t_i,f_i} = 0$. When $x^u_{t_i,f_i} = 0$, it leads to no power

assignment to the RB and at most P_{max} of power is assigned to RB if $x^u_{t_i,f_i} = 1$.

- *Step 2 (Probabilistic to Non-Probabilistic Transformation):* The probabilistic constraint (C4) is a cause for the non-convexity of (P1a). To address this issue, the probabilistic constraint needs to be transformed into the non-probabilistic constraint. In this context, we use the following approximation [34], [39] to transform the probabilistic constraint into a non-probabilistic as

$$\begin{aligned}\Pr\left(\mathcal{C}^u_{t_i,f_i} < \mathcal{D}^u_{t_i,f_i}\right) &\overset{(a)}{=} \Pr\left(\varphi^u_{t_i,f_i} < \gamma^u_{t_i,f_i}\right) \\ &= \Pr\left[\left|h^u_{t_i,f_i}\right|^2 p^u_{t_i,f_i} d^{-v}_{BS,u} < S^u_{t_i,f_i}\right] \\ &= \Pr\left[\left|h^u_{t_i,f_i}\right|^2 < \frac{S^u_{t_i,f_i}}{p^u_{t_i,f_i} d^{-v}_{BS,u}}\right] \\ &= F_{\left|h^u_{t_i,f_i}\right|^2}\left(\frac{S^u_{t_i,f_i}}{p^u_{t_i,f_i} d^{-v}_{BS,u}}\right) \end{aligned} \tag{16}$$

(a) gives that the outage probability is bounded by $\Pr(\varphi^u_{t_i,f_i} < \gamma^u_{t_i,f_i})$ [39]. Since $\Pr(\mathcal{C}^u_{t_i,f_i} < \mathcal{D}^u_{t_i,f_i}) \leq \Delta_{out}$,

$$S^u_{t_i,f_i} = F^{-1}_{\left|h^u_{t_i,f_i}\right|^2}(\Delta_{out})\; p^u_{t_i,f_i} d^{-v}_{BS,u} \tag{17}$$

where $F^{-1}_{|h^u_{t_i,f_i}|^2}(\cdot)$ is the inverse cumulative distribution function (CDF) of a non-central chi-square with non-centrality parameter $\frac{2|\hat{h}^u_{t_i,f_i}|^2}{\sigma_e^2}$ and 2 degrees of freedom [34]. By using (17), the achievable SNR of user 'u' on the RB with the estimated CSI is written as

$$\tilde{\gamma}^u_{t_i,f_i} = \frac{F^{-1}_{\left|h^u_{t_i,f_i}\right|^2}(\Delta_{out}) p^u_{t_i,f_i} d^{-v}_{BS,u}}{\sigma_o^2} \tag{18}$$

By using (18), it is possible to incorporate the outage constraint (C4) into the minimum SNR constraint (C5). Now, the power-minimization problem (P1a) is reformulated as

$$\begin{aligned}(\mathbf{P1b:}) \quad &\min_{\left\{x^u_{t_i,f_i},\; p^u_{t_i,f_i}\right\}} \quad \hat{\boldsymbol{\mathcal{P}}}_{\boldsymbol{tot}} \\ &\text{subject to} \\ &(C1) - (C3) \quad \text{in } (P1a) \\ &(C5)\text{: } \tilde{\gamma}^u_{t_i,f_i} \geq x^u_{t_i,f_i}\gamma^s_{th}, \forall t_i, f_i, u \\ &(C6) - (C10) \quad \text{in } (P1a) \end{aligned} \tag{19}$$

- *Step 3 (Penalized Formulation):* The optimization problem (**P1b**) is combinatorial due to the binary constraint (C1). To circumvent the combinatorial nature of the optimization problem (**P1b**), we relax the binary variable $x^u_{t_i,f_i}$ between 0 and 1; this relaxation renders (**P1b**) as a convex problem. However, the relaxed problem is not guaranteed to obtain the binary solutions for $x^u_{t_i,f_i}$s. In the context resource allocation attaining strict binary solutions for $x^u_{t_i,f_i}$ is paramount. Therefore, to promote the binary $x^u_{t_i,f_i}$s, the relaxed problem is further penalized as follows

$$\begin{aligned}(\mathbf{P1c:}) \quad &\min_{\left\{x^u_{t_i,f_i},\; p^u_{t_i,f_i}\right\}} \quad \hat{\boldsymbol{\mathcal{P}}}_{\boldsymbol{tot}} - \lambda_1 \sum_i \sum_{t_i} \sum_{f_i} \sum_u \mathbf{Y}\left(x^u_{t_i,f_i}\right) \\ &\text{subject to} \\ &(C1)\text{: } 0 \leq x^u_{t_i,f_i} \leq 1; \forall t_i, f_i, u \\ &(C2), (C3) \text{ and } (C5 - C10) \quad \text{in } (P1b) \end{aligned} \tag{20}$$

where λ_1 is a penalty factor and $\mathbf{Y}(x^u_{t_i,f_i})$ is penalty function for promoting binary solutions. We consider $\mathbf{Y}(x^u_{t_i,f_i}) = (x^u_{t_i,f_i})^2 - x^u_{t_i,f_i}$ as a penalty function [40], which is convex in the region of [0,1]. This function induces no penalty at the value of $x^u_{t_i,f_i}$ is either '0' or '1', and increases the penalty as the value of $x^u_{t_i,f_i}$ moves away from '0' or '1' with the maximum penalty at $x^u_{t_i,f_i} = 0.5$. With appropriate value for λ_1, the binary nature of $x^u_{t_i,f_i}$s in (**P1c**) can be ensured.

- *Step 4 (DC Programming):* Now, the problem (**P1c**) is a class of DC problem, since the objective function of (**P1c**) is a difference of two convex functions and constraint are linear and convex. In this regard, we utilize the successive convex approximation (SCA) algorithm to the DC problem in (20). In this algorithm, the following two steps are performed iteratively until the convergence:

(i) *Convexification by Taylor series approximation:* Assume $(x^u_{t_i,f_i})^{j-1}$ is the estimate of $x^u_{t_i,f_i}$ in the $(j-1)^{th}$ iteration. In the j^{th} iteration, the concave part of the objective in **P1c**, i.e., $-\sum_i \sum_{t_i} \sum_{f_i} \sum_u \mathbf{Y}(x^u_{t_i,f_i})$ is replaced by its first order Taylor approximation around the estimate of $(x^u_{t_i,f_i})^{j-1}$

$$\begin{aligned}-\mathbf{Y}\left(x^u_{t_i,f_i}\right) \leq &-Y\left(\left(x^u_{t_i,f_i}\right)^{j-1}\right) \\ &- \left(\left(x^u_{t_i,f_i}\right) - \left(x^u_{t_i,f_i}\right)^{j-1}\right)\nabla\mathbf{Y}\left(x^u_{t_i,f_i}\right) \end{aligned} \tag{21}$$

Algorithm 1 Successive Convex Approximation for Resource Allocation Problem

1: **Initialization:** Tolerance: $\mu = 10^{-5}$, $\Upsilon = 1$, $\mathcal{P}^{(0)}_{tot} = 0$, penalty parameters: λ_1, λ_2, $j = 1$, and initial point : $(x^u_{t_i,f_i})^{(0)}$
2: **while** $\Upsilon \geq \mu$ **do**
3: (**i**) Obtain $\mathcal{P}^{(j)}_{tot}$ and $\{(x^u_{t_i,f_i})^j, (p^u_{t_i,f_i})^j\}$ by solving (**P1d**) for the slice-isolation resource allocation problem.
4: (**ii**) Obtain $\mathcal{P}^{(j)}_{tot}$ and $\{(x^u_{t_i,f_i})^j, (p^u_{t_i,f_i})^j\}$ by solving (**P2d**) for the slice-aware resource allocation problem.
5: **Update:** $\Upsilon = |\mathcal{P}^{(j)}_{tot} - \mathcal{P}^{(j-1)}_{tot}|$
6: Set $j = j + 1$
7: **end while**
8: **Output:** $(x^u_{t_i,f_i})^* = (x^u_{t_i,f_i})^j$
9: $(p^u_{t_i,f_i})^* = (p^u_{t_i,f_i})^j$ for $(x^u_{t_i,f_i})^j = 1$, and else 0.

(ii) *Optimization:* The next update $(x^u_{t_i,f_i})^{j+1}$ is acquired by solving the convex problem given in **P1d**. Notice that the objective in (**P1d**) is an upper bound to **P1c**. Hence, the problem (**P1d**) is an upper bound to the problem in (**P1c**) where the bound is tight at the current iteration.

Feasible initial point (FIP) selection: Since a possible initial point is essential for the SCA-based algorithm's efficient performance, we obtain a feasible initial point (FIP) that converges to a stationary point of the problem **P1c** using the following steps:

Step 1: Allocate the equal transmit powers to all the available RBs (i.e., $p^u_{t_i,f_i} = P_{max}/\sum_{i=1}^{3} F_i$).

Step 2: Solve the following optimization problem

$$\begin{aligned}(\textbf{FIP1:}) \quad & \min_{\left\{x^u_{t_i,f_i}\right\}} \hat{\mathcal{P}}_{tot} \\ & \text{subject to} \\ & (C1):\ 0 \le x^u_{t_i,f_i} \le 1;\ \forall t_i, f_i, u \\ & (C2), (C3), \text{and} (C5) - (C8) \ \text{ in } (P1b) \end{aligned} \tag{23}$$

By solving the above optimization problem, we obtain the FIP for solving the strict slice-isolation problem. Note that the FIP need not to be binary, but it must satisfy all the other optimization problem constraints.

B. SOLUTION TO SLICE-AWARE PROBLEM

The similar approach as of to solve the problem (P1) is also applied to solve the problem (P2). By following the similar procedures given in Step 1-Step 3 of the slice-isolation solution, the optimization problem for slice-aware approach is reformulated as

$$\begin{aligned}(\textbf{P2c:}) \quad & \min_{\left\{x^u_{t_i,f_i},\ p^u_{t_i,f_i}\right\}} \hat{\mathcal{P}}_{tot} - \lambda_2 \sum_i \sum_{t_i} \sum_{f_i} \sum_u \mathbf{Y}\left(x^u_{t_i,f_i}\right) \\ & \text{subject to} \\ & (C1):\ 0 \le x^u_{t_i,f_i} \le 1;\ \forall t_i, f_i, u \\ & (C2), (C5), (C6) - (C10) \ \text{in } (\textbf{P1b}), \\ & (C11) - (C13) \ \text{in } (\textbf{P2}) \end{aligned} \tag{24}$$

To solve the problem (**P2c**), we need to apply the steps (i) Taylor series approximation and (ii) update of feasible point, which is done by solving the problem (**P2d**).

In order to solve the problem (P2d), we need an FIP that can be obtained by using the following two steps:

Step 1: Allocate the equal transmit powers to all the available RBs (i.e., $p^u_{t_i,f_i} = P_{max}/\sum_{i=1}^{3} F_i$).

Step 2: Solve the following optimization problem

$$\begin{aligned}(\textbf{FIP2:}) \quad & \min_{\left\{x^u_{t_i,f_i}\right\}} \hat{\mathcal{P}}_{tot} \\ & \text{subject to} \\ & (C1):\ 0 \le x^u_{t_i,f_i} \le 1;\ \forall t_i, f_i, u \\ & (C5) \ \text{in } (P1b) \\ & (C2), (C3), (C6) - (C8), (C11) - (C13) \ \text{in } (P2) \end{aligned} \tag{26}$$

Optimization problems in (22) and (25), as shown at the bottom of the page, are convex because objective functions are convex, and constraints are linear and convex. Thus, these problems can be solved easily by using standard convex optimization tools like CVX [41]. Algorithm 1 briefs the steps of an iterative procedure involved in solving problems (**P1c**) and (**P2c**). This algorithm provides the improved feasible solution in every iteration until convergence to local optimum points of problems (**P1c**) and (**P2c**) in polynomial time. Further, this algorithm converges to a stationary point typically within $5-7$ iterations.

C. COMPUTATIONAL COMPLEXITY ANALYSIS

Since the proposed scheduling algorithm is an SCA based iterative algorithm, its complexity depends on the complexities of following two procedures: (i) the convex-concave procedure involved in the convex problem (P1d) or (P2d), and (ii) the procedure for the selection of the initial feasible point using the convex problem (FIP1) or (FIP2).

Assume that variables U_i, and U represent the users associated with numerology 'i', and total number users in the network, respectively.

1) COMPLEXITY OF STRICT SLICE-ISOLATION SCHEME

- *With mixed-numerologies in the frequency domain:* The convex problem (P1d) has $2(L\mathcal{T}_1\mathcal{F}_1 + K\mathcal{T}_2\mathcal{F}_2 + M\mathcal{T}_3\mathcal{F}_3)$ decision variables, $5(L\mathcal{T}_1\mathcal{F}_1 + K\mathcal{T}_2\mathcal{F}_2 + M\mathcal{T}_3\mathcal{F}_3) + (\mathcal{T}_1\mathcal{F}_1 + \mathcal{T}_2\mathcal{F}_2 + \mathcal{T}_3\mathcal{F}_3) + U$ linear constraints, and $\mathcal{T}_1 + \mathcal{T}_2 + \mathcal{T}_3$ convex constraints. Thus, the computational complexity of the convex problem (P1d) is $\mathcal{O}((2(L\mathcal{T}_1\mathcal{F}_1 + K\mathcal{T}_2\mathcal{F}_2 + M\mathcal{T}_3\mathcal{F}_3))^3(\mathcal{T}_1(\mathcal{F}_1(5L+1)+1)+ \mathcal{T}_2(\mathcal{F}_2(5K+1)+1)+\mathcal{T}_3(\mathcal{F}_3(5M+1)+1)+U))$ [42]. The

$$\begin{aligned}(\textbf{P1d:}) \quad & \min_{\left\{x^u_{t_i,f_i},\ p^u_{t_i,f_i}\right\}} \hat{\mathcal{P}}_{tot} - \lambda_1 \sum_i \sum_{t_i} \sum_{f_i} \sum_u \left(\left(x^u_{t_i,f_i}\right) - \left(x^u_{t_i,f_i}\right)^{j-1}\right) \nabla \mathbf{Y}\left(x^u_{t_i,f_i}\right) \\ & \text{subject to} \ \ (C1) - (C3), (C5 - C10) \ \text{in } (\textbf{P1b}) \end{aligned} \tag{22}$$

$$\begin{aligned}(\textbf{P2d:}) \quad & \min_{\left\{x^u_{t_i,f_i},\ p^u_{t_i,f_i}\right\}} \hat{\mathcal{P}}_{tot} - \lambda_2 \sum_i \sum_{t_i} \sum_{f_i} \sum_u \left(\left(x^u_{t_i,f_i}\right) - \left(x^u_{t_i,f_i}\right)^{j-1}\right) \nabla \mathbf{Y}\left(x^u_{t_i,f_i}\right) \\ & \text{subject to} \ \ (C1), (C2), (C5 - C13) \ \text{in } (\textbf{P2c}) \end{aligned} \tag{25}$$

convex problem (FIP1) has $(L\mathcal{T}_1\mathcal{F}_1 + K\mathcal{T}_2\mathcal{F}_2 + M\mathcal{T}_3\mathcal{F}_3)$ decision variables, and $3(L\mathcal{T}_1\mathcal{F}_1 + K\mathcal{T}_2\mathcal{F}_2 + M\mathcal{T}_3\mathcal{F}_3) + (\mathcal{T}_1\mathcal{F}_1 + \mathcal{T}_2\mathcal{F}_2 + \mathcal{T}_3\mathcal{F}_3) + U$ linear constraints. Thus, the computational complexity of the convex problem (FIP1) is $\mathcal{O}(((L\mathcal{T}_1\mathcal{F}_1 + K\mathcal{T}_2\mathcal{F}_2 + M\mathcal{T}_3\mathcal{F}_3))^3(\mathcal{T}_1(\mathcal{F}_1(3L+1)) + \mathcal{T}_2(\mathcal{F}_2(3K+1)) + \mathcal{T}_3(\mathcal{F}_3(3M+1)) + U))$.

- *With mixed-numerologies in the time domain:* The convex problem (P1d) has $2(U_i\mathcal{T}_i\mathcal{F}_i)$ decision variables, $5(U_i\mathcal{T}_i\mathcal{F}_i) + (\mathcal{T}_i\mathcal{F}_i) + U_i$ linear constraints, and $(\mathcal{T}_i)$ convex constraints. Thus, the computational complexity of the convex problem (P1d) is $O((2(U_i\mathcal{T}_i\mathcal{F}_i))^3(\mathcal{T}_i\mathcal{F}_i(5U_i + 1) + U_i))$. The convex problem (FIP1) has $2(U_i\mathcal{T}_i\mathcal{F}_i)$ decision variables, $3(U_i\mathcal{T}_i\mathcal{F}_i) + (\mathcal{T}_i\mathcal{F}_i) + U_i$ linear constraints. Thus, the computational complexity of the convex problem (FIP1) is $O((2(U_i\mathcal{T}_i\mathcal{F}_i))^3(\mathcal{T}_i\mathcal{F}_i(3U_i + 1) + U_i))$.

2) COMPLEXITY OF SLICE-AWARE SCHEME

- *With mixed-numerologies in the frequency domain:* The convex problem (P2d) has $2U(\mathcal{T}_1\mathcal{F}_1 + \mathcal{T}_2\mathcal{F}_2 + \mathcal{T}_3\mathcal{F}_3)$ decision variables, $(5U+1)(\mathcal{T}_1\mathcal{F}_1 + \mathcal{T}_2\mathcal{F}_2 + \mathcal{T}_3\mathcal{F}_3) + 3U$ linear constraints, and $(\mathcal{T}_1 + \mathcal{T}_2 + \mathcal{T}_3)$ convex constraints. Thus, the computational complexity of the convex problem (P2d) is $\mathcal{O}((2U(\mathcal{T}_1\mathcal{F}_1 + \mathcal{T}_2\mathcal{F}_2 + \mathcal{T}_3\mathcal{F}_3))^3((5U+1)(\mathcal{T}_1\mathcal{F}_1 + \mathcal{T}_2\mathcal{F}_2 + \mathcal{T}_3\mathcal{F}_3) + 3U + \sum_i^3 \mathcal{T}_i))$. The convex problem (FIP2) has $U(\mathcal{T}_1\mathcal{F}_1 + \mathcal{T}_2\mathcal{F}_2 + \mathcal{T}_3\mathcal{F}_3)$ decision variables, and $(3U+1)(\mathcal{T}_1\mathcal{F}_1 + \mathcal{T}_2\mathcal{F}_2 + \mathcal{T}_3\mathcal{F}_3) + 3U$ linear constraints. Thus, the computational complexity of the convex problem (FIP2) is $\mathcal{O}(((U(\mathcal{T}_1\mathcal{F}_1 + \mathcal{T}_2\mathcal{F}_2 + \mathcal{T}_3\mathcal{F}_3)))^3((3U+1)(\mathcal{T}_1\mathcal{F}_1 + \mathcal{T}_2\mathcal{F}_2 + \mathcal{T}_3\mathcal{F}_3) + 3U))$.
- *With mixed-numerologies in the time domain:* The convex problem (P2d) has $2(U\mathcal{T}_i\mathcal{F}_i)$ decision variables, $5(U\mathcal{T}_i\mathcal{F}_i) + (\mathcal{T}_i\mathcal{F}_i) + U$ linear constraints, and $(\mathcal{T}_i)$ convex constraints. Thus, the computational complexity of the convex problem (P2d) is $\mathcal{O}((2(U\mathcal{T}_i\mathcal{F}_i))^3(\mathcal{T}_i\mathcal{F}_i(5U+1) + U))$. The convex problem (FIP2) has $2(U\mathcal{T}_i\mathcal{F}_i)$ decision variables, $3(U\mathcal{T}_i\mathcal{F}_i) + (\mathcal{T}_i\mathcal{F}_i) + U$ linear constraints. Thus, the computational complexity of the convex problem (FIP2) is $\mathcal{O}((2(U\mathcal{T}_i\mathcal{F}_i))^3(\mathcal{T}_i\mathcal{F}_i(3U+1) + U))$.

From the complexity analysis, it is evident that the slice-isolation scheme's complexity is lower than the slice-aware scheme. In particular, the proposed algorithm's complexity with mixed-numerologies in the time domain is lower than the others. Also, the computational complexity of proposed algorithms is higher for the larger dimensions. However, the upcoming 5G systems foresee to utilize the cloud-computing technology, which can efficiently handle the higher computational complexities, including the proposed resource-scheduling algorithm's complexity.

V. NUMERICAL EVALUATIONS

In this section, simulation results are provided to compare the performances of the proposed joint allocation of power and RBs technique for both slice-isolation and slice-aware radio resource scheduling mechanisms, and with fixed and mixed-numerologies to the active eMBB, URLLC and mMTC users in the OFDMA based DL wireless network. Importantly, we compare the performance of the proposed scheduling techniques in terms of achievable data rates, latency in delivered packets, and queue status (i.e., key performance indicators (KPIs)) for eMBB, URLLC, and mMTC users, respectively.

A. SIMULATION ENVIRONMENT

In the considered DL wireless network, a BS is located at the center of the cell coverage area with the radius of 250m and UEs from different services are uniformly distributed across the total coverage area of the cell. In this scenario, we consider the small scale Raleigh fading channel coefficients (i.e., drawn from $\mathcal{CN}(0, 1-\sigma_e^2)$, $\sigma_e^2 = 0.1$ or 0.01) between the BS and users and the 3GPP urban path loss model with the path loss exponent of 3.76. Further, the channel estimation errors at the BS have entries drawn from $\mathcal{CN}(0, \sigma_e^2)$. The total simulation time considers 10 sub-frames (i.e., 10 msec), where the radio resource optimization is executed at every sub-frame (i.e., 18MHz $\times$ 1msec) taking into account the actual user queue status. Furthermore, mini-slot based frame structure is considered where each slot comprises of 7 OFDM symbols.

Mixed numerologies in frequency domain: The complete transmission bandwidth (i.e., 18 MHz) is divided into 58 sub-bands, where 34 sub-bands have frame Numerology-1 ($\mu = 0$), 16 sub-bands have Numerology-2 ($\mu = 1$) and 8 sub-bands have Numerology-3 ($\mu = 2$). Moreover, the consecutive sub-bands with different numerologies are separated by a guard-band of 180 KHz. Therefore, the RB grid within a sub-frame is composed of 34×2 RBs for $\mu = 0$, 16×4 RBs for $\mu = 1$, and 8×8 RBs for $\mu = 2$.

Mixed numerologies in time domain: In this case, each sub-frame of 1 msec duration operates under a specific numerology. For Numerology-1 ($\mu = 0$), the RB grid within a sub-frame is composed of 100×2 RBs; for Numerology-2 ($\mu = 1$), the RB grid within a sub-frame is composed of 50×4 RBs; and for Numerology-3 ($\mu = 2$), the RB grid within a sub-frame is composed of 25×8 RBs.

Fixed numerology: In this model, all the available RBs utilize the same Numerology-1 ($\mu = 0$). Therefore, the RB grid within a sub-frame is composed of 100×2 RBs.

The complete set of the simulation parameters are summarized in Table 4.

B. RESULTS AND DISCUSSIONS

We initially assume a scenario with $L = 5$ URLLC users, $M = 15$ mMTC users and $K = \{5, 10\}$ eMBB users and we evaluate the achievable user data rate of eMBB users. In Fig. 4 and Fig. 5, we show the empirical cumulative distribution function (ECDF) of achievable transmission data rates of eMBB users on every sub-frame (i.e., time-duration of $1ms$) by using the considered slice-isolation and slice-aware resource scheduling mechanisms for allocating the time-frequency radio resources with mixed-numerologies in frequency and time domains. In particular, Fig. 4 focuses on the mixed numerologies in frequency domain, while Fig. 5

TABLE 4. Simulation parameters.

Parameter	Value
Cell radius	$250m$
Number of eMBB users (K)	$5, 10$
Number of URLLC users (L)	$5, 10$
Number of mMTC users (M)	$15, 20, 25$
BS Max. Transmit power (P_{max})	50 dBm
Path loss exponent (ν)	3.76
Channel	Rayleigh fading model
CSI error variance (σ_e^2)	$0.01, 0.1$
Channel outage probability (Δ_{out})	$0.1, 0.3$
Total simulation time	10 msec (10 sub-frames)
Effective sub-frame bandwidth	18 MHz
Sub-frame duration	1 msec
RB grid size for fixed numerology	100×2 RBs
RB grid size for mixed numerology in frequency domain	34×2 RBs for $\mu = 0$, 16×4 RBs for $\mu = 1$, and 8×8 RBs for $\mu = 2$
RB grid size for mixed numerology in time domain	100×2 RBs for $\mu = 0$; 50×4 RBs for $\mu = 1$; and 25×8 RBs for $\mu = 2$
Number of sub-carriers per RB	12
Number of OFDM symbols per RB	7
Number of REs per RB	84
REs within RB devoted to reference signals	24
Guard band	1 RB of 180 KHz
Traffic model for URLLC	FTP3 model
URLLC packet size (B)	32 bytes
Traffic model for eMBB	Full-buffered
eMBB packet size	Infinite
Traffic model for mMTC	Pareto distribution
mMTC packet size	20 bytes to 200 bytes

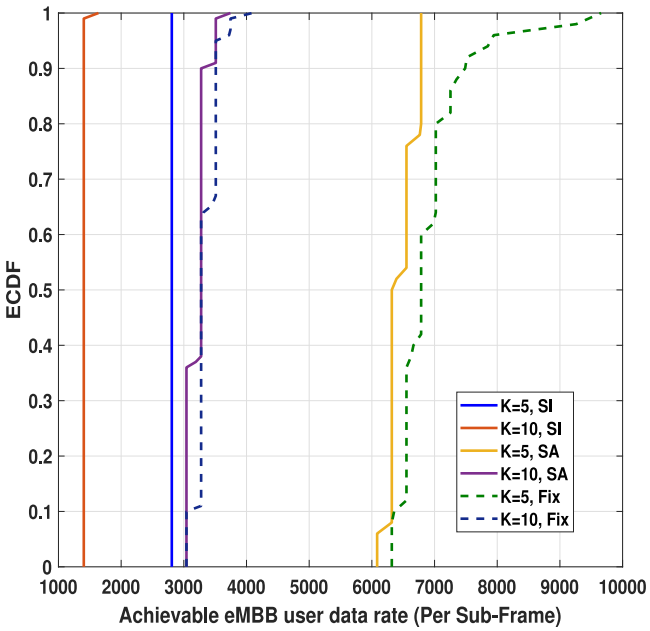


FIGURE 4. CDF of achievable data rates for eMBB users using slice-isolation and slice-aware resource allocation mechanisms with mixed numerologies in frequency domain and fixed numerology. ($L = 5$, $M = 15$, $\sigma_e^2 = 0.01$, $\Delta_{out} = 0.1$, $\rho = 1$, $\kappa = \xi u$).

focuses on the mixed numerologies in time domain. Also, for the comparison, we show the performance of the resource scheduling scheme with the conventional fixed-numerology based time-frequency resources grid model.

As expected, the results in Fig. 4 and Fig. 5 confirm that the slice-aware resource allocation scheme outperforms

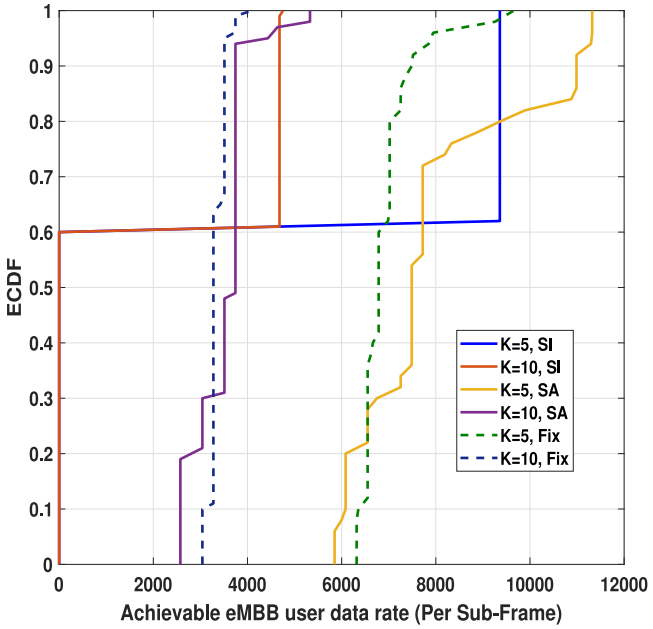


FIGURE 5. CDF of achievable data rates for eMBB users using slice-isolation and slice-aware resource allocation mechanisms with mixed numerologies in time domain and fixed numerology. ($L = 5$, $M = 15$, $\sigma_e^2 = 0.01$, $\Delta_{out} = 0.1$, $\rho = 1$, $\kappa = \xi u$).

the slice-isolation based resource allocation scheme in terms of the achievable eMBB data rates. This is because the slice-aware resource allocation scheme first allocates the resources to the eMBB users from its dedicated resource slice. Later, it assigns some of the underutilized resources to the eMBB users from the other slices. On the contrary, the slice-isolation scheme only assigns the resources to the eMBB users from its dedicated resource slice. As a result, the eMBB users cannot achieve the higher data transmission rate.

From the results in Fig. 4, we observe that allocating the RBs of mixed-numerologies in the frequency domain using the slice-aware resource scheduling scheme achieves almost the same rate performance as of the performance of fixed-numerology for eMBB users. This was expected, as the fixed-numerology scheme was designed particularly for the eMBB service. As shown later on in this article, fixed-numerology fails in satisfying other services requirements.

Also, we observe that by allocating the radio resources using the slice-isolation resource scheduling algorithm, every scheduled eMBB user achieves almost the same transmission data rate on every sub-frame. Through the slice-isolation scheduling scheme, eMBB service receives the dedicated number of radio resources that are equally allocated to the scheduled eMBB users to maintain the fairness between users in the allocation of resources. Also, every RB is scheduled to the user with optimal power to obtain a certain SNR threshold. Because of these two factors, every scheduled eMBB user achieve similar data rate on every sub-frame. Further, as expected, it is observed that the achievable rate of each eMBB user reduces when the number of active

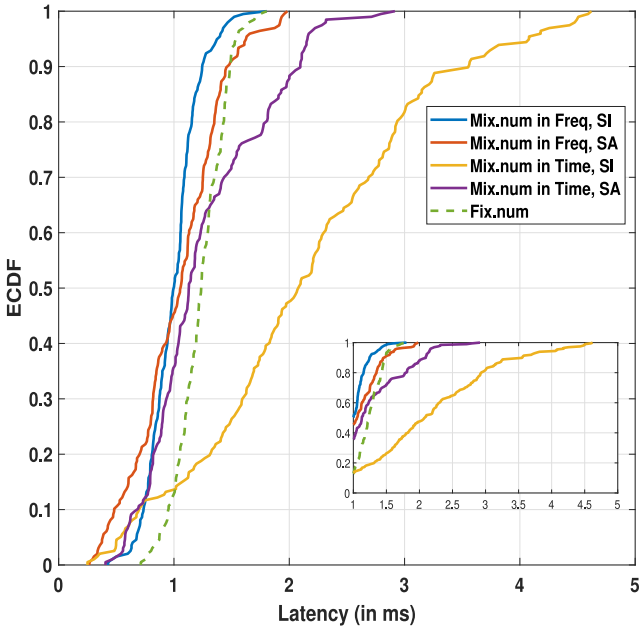


FIGURE 6. CDF of achieved latency in delivered URLLC packets using the slice-isolation and slice-ware scheduling algorithms with the different frame numerology models. ($K = 5$, $L = 5$, $M = 15$, $\lambda_U = 4$ packets/1ms, $\sigma_e^2 = 0.01$, $\Delta_{out} = 0.1$, $\rho = 1$, $\kappa = \xi_U$).

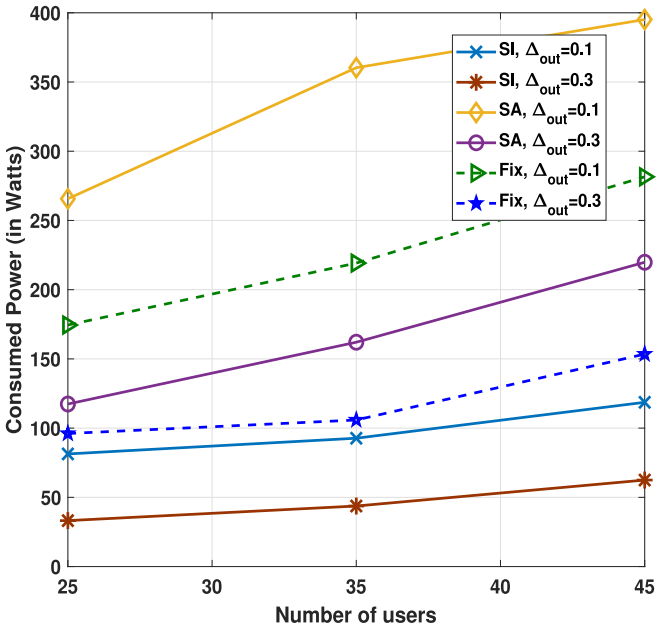


FIGURE 7. Consumed power vs the number of users with varying channel outage probabilities using the SI and SA scheduling algorithms with the mixed-numerology in time. The CSI estimation error is set to $\sigma_e^2 = 0.01$.

eMBB users in cell increases. This is justified by the fact that resources need to be distributed across a higher number of users.

Similar conclusions are extracted for the mixed numerologies in the time domain results shown in Fig. 5. It is evident that by performing the slice-isolation scheduling scheme for allocation of RBs with the mixed-numerologies in time, the eMBB users cannot obtain the minimum data rate on most of the sub-frames. With the mixed-numerologies in time, the slice-isolation scheme assigns every sub-frame of 1ms time duration to a specific service with unique numerology. Therefore, the eMBB users cannot receive the resources from the slices assigned to the other services. In contrast to this case, by using the slice-aware scheme, the eMBB users achieve data rates on every sub-frame through accessing the underutilized resource of other slices.

In particular, slice-aware scheme with the mixed-numerologies in time achieves a slightly better performance in terms of data rate for eMBB users compared to the mixed-numerologies in frequency and fixed numerology. With the mixed-numerologies in frequency domain, some resource blocks are wasted for the guard bands between two consecutive numerologies. This is one of the reasons to receive the lower eMBB rates with mixed numerologies in the frequency domain. However, this gain comes at a cost of complexity in synchronization, as discussed in Section II.

Next, we evaluate the performance of the proposed technique in terms of URLLC QoS satisfaction. In particular, we illustrate the ECDF of achievable latency in the delivered URLLC packets in Fig. 6, by performing the considered slice-isolation and slice-aware resource scheduling schemes with different time-frequency RB grids. Herein, the latency is computed as the gap of the time that the packet has entered the queue and the time that the packet has been scheduled and vacated from the queue. Thus, the complete packet latency is estimated as the accumulation of the packet waiting time in the queue, the time required for the RB assignment, and the data transmission delay. From Fig. 6, it is clear that considering mixed-numerologies in frequency domain, the proposed algorithm provides lower URLLC latencies as compared to the fixed numerology. For the considered traffic arrival rate $\lambda_U = 4$, the mixed numerologies in frequency domain achieve the best performance, with similar latencies with both slice-aware and slice-isolation radio resource allocation mechanisms. This evidences that for relatively low traffic loads, the continuous use of high numerologies (i.e., $\mu = 2$) of the frequency domain mixed-numerologies is preferable to the intermittent user of numerologies of the time mixed-numerologies.

In order to show the impact of the imperfect CSI on the overall consumed power of the proposed slicing based resource scheduling scheme, we evaluate the effect of the imperfect CSI using the channel outage probability.

Fig. 7 and Fig. 8 illustrate the power consumption of the BS versus the number of active users in the cell with the different channel outage probability levels, for the mixed numerologies in time and frequency domains, respectively. Focusing on Fig. 7, the power consumption of the BS increases when the channel outage probability decreases. For instance, when the number of active users in the cell is 35, the consumed power of the BS using the fixed-numerology based resource scheduling scheme with $\Delta_{out} = 0.1$ is 52% more than that with $\Delta_{out} = 0.3$. This was expected, as the

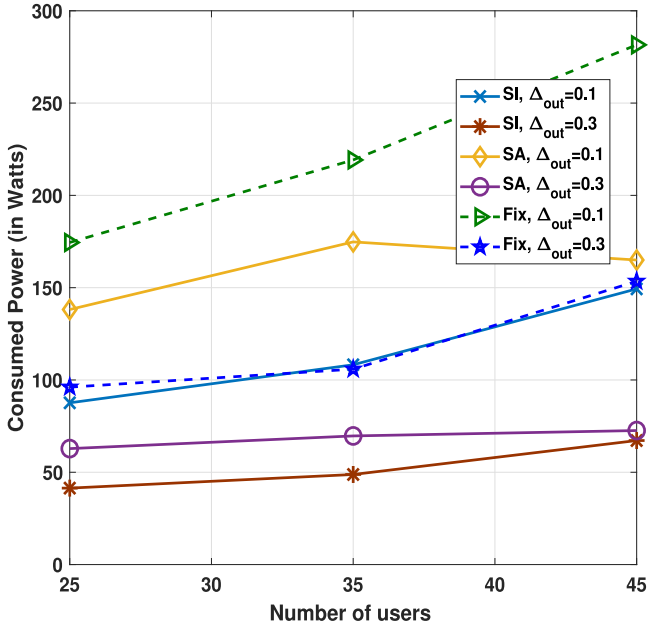


FIGURE 8. Consumed power vs the number of users with the mixed-numerology in Frequency under different channel outage probabilities. The CSI estimation error is set to $\sigma_e^2 = 0.01$.

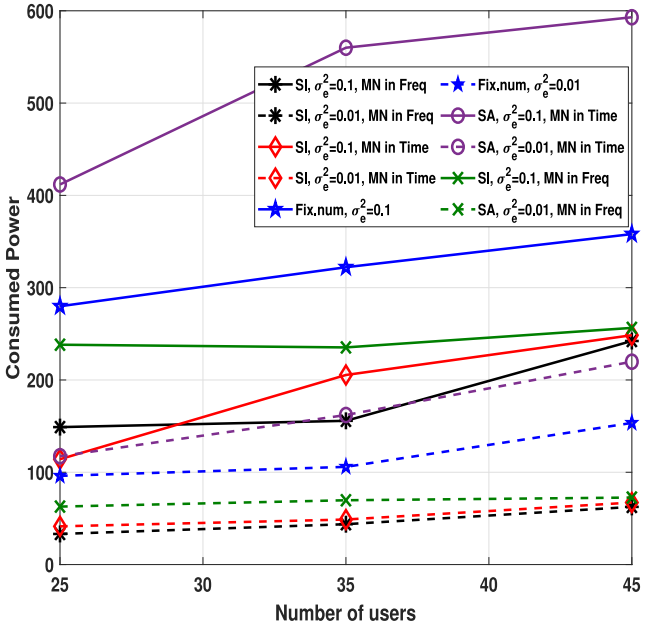


FIGURE 9. Consumed power vs number of users with varying CSI estimation error variances. The channel outage probability is set to $\Delta_{out} = 0.3$.

system requires a larger amount of power to achieve better channel outage probability. This power consumption difference is more pronounced by using the slice-aware resource scheduling algorithm with the mixed-numerologies. By comparing the curves of slice-aware scheme with $\Delta_{out} = 0.1$ and $\Delta_{out} = 0.3$, it can be observed that with $\Delta_{out} = 0.1$ the scheme consumes 2 times more power than that with $\Delta_{out} = 0.3$. The required transmit power also increases with the total number of users in the system.

Also, the results in Fig. 7 and Fig. 8 show that the slice-isolation scheme consumes less power than the slice-aware scheme. Due to the allocation of fewer RBs to the users, the slice-isolation scheme consumes less power than the slice-aware scheme. In particular, by comparing the results in Fig. 7 and Fig. 8 illustrate that the slice-aware resource allocation scheme with mixed-numerology in time consumes more power than that with the mixed-numerology in frequency. With the mixed-numerology in the time domain, the slice-aware scheme allocates the RBs of every sub-frame to the active users from different services. In this scenario, some users may get the RBs with poor channel conditions that consume more power to achieve the particular QoS requirements. On the contrary, with mixed-numerologies in the frequency domain, the slice-aware scheme allocates the RBs to the users from its dedicated resource slice according to their requesting service. Later, it assigns the underutilized RBs to the users from other slices if required. Mostly, due to less traffic load, the URLLC and mMTC users cannot access the RBs from other slices, while eMBB users take the underutilized RBs from other slices because of their data rate requirement. Therefore, most users receive RBs with the best channel conditions that consume less power to satisfy their QoS requirements. Further, different from mixed-numerologies in the time domain, with mixed-numerologies in the frequency domain, some of the resources are wasted for guards, which is also one reason for less power consumption.

Also, it is clear from the results in Fig. 7 that the algorithm with fixed-numerology shows less power consumption than the slice-aware scheme with mixed-numerologies in the time-domain. With mixed-numerologies in the time-domain, the available number of RBs is varied (i.e., 25, 50, and 100) according to the chosen numerology. On the contrary, with the fixed numerology (i.e., with SCS 15kHz), the same number of RBs (i.e., 100) is available for allocation in every time-slot. Therefore, more users get right channel conditions on scheduled RBs due to the availability of a large set of RBs with fixed numerology that leads to less power consumption than mixed-numerologies in the time-domain.

In Fig. 9, we evaluate the performance of the proposed algorithms in the power consumption of BS versus the number of users with different CSI estimation error variances. From the results in Fig. 9, it is noticed that by increasing the CSI error variance, the power consumption of the system increases. For example, when the number of active users in the cell is 35, by using the slice-aware resource allocation, the power consumption of the BS with $\sigma_e^2 = 0.01$ is almost 3 times lesser than that with $\sigma_e^2 = 0.1$. From these results, it is confirmed that the channel estimation error can degrade scheduling algorithms' performance in terms of power consumption.

Fig. 10 illustrates the accumulation of the queue status (i.e., unscheduled packets) of mMTC users on every sub-frame after executing the scheduling process using the proposed algorithms with mixed-numerologies and fixed

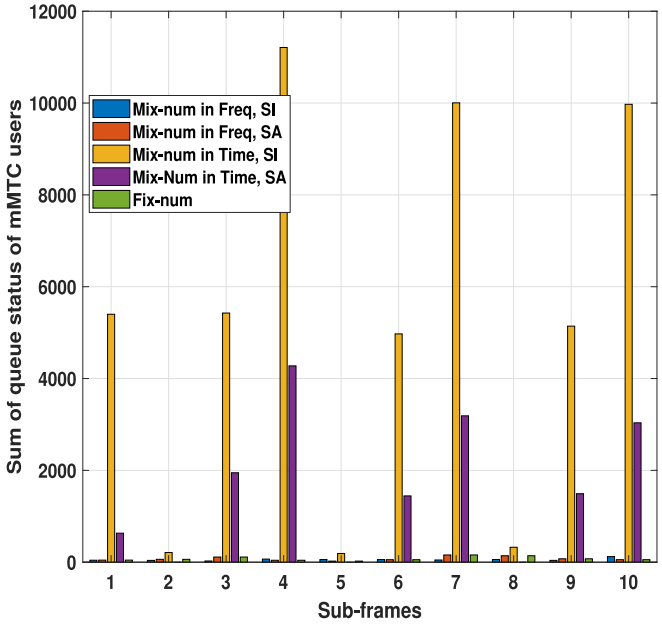


FIGURE 10. Queue status of mMTC users with mixed numerologies in frequency, time domains, and with fixed-numerology.

numerology. As can be seen from the results in Fig. 10, it is clear that with mixed-numerologies in frequency and fixed numerology, the scheduling algorithms allocate the resources to vacate the mMTC queues. But, with mixed numerologies in time, the slice-isolation scheme fails to provide the resources to mMTC users on every sub-frame while slice-aware provides some resources to mMTC user on every sub-frame to vacate the mMTC queues. Further, by tuning the control parameter (i.e., ρ), the slice-aware scheduler may allocate the more number of resources for mMTC service to vacate the complete queues at the cost of eMBB rates.

The simulation results in Fig. 4 to Fig. 10 show the trade off between the power consumption and performance of the users from different services. Using the slice-aware scheduling scheme, the eMBB users achieve the higher data rates, the URLLC users achieve the better performance in terms of latency and mMTC queues are vacated while consuming the more power at the BS. In contrast, slice-isolation based scheduling scheme consumes the less power, but fails to achieve the higher data rates for eMBB users and also fails to achieve low latencies for URLLC users, and unable vacate the mMTC queues on every sub-frame with the mixed-nuemrologies in time. In particular, the slice-aware scheme with the mixed-numerologies in frequency shows the approximately same performance as of the other schemes in terms of eMBB rates. It outperforms the other slice-aware scheme with mixed numerologies in time and fixed-numerology in terms of latency in URLLC packets' delivery. It also shows the same performance as the performance of slice-isolation with mixed-numerologies in frequency, fixed numerology. It outperforms the scheduling schemes with mixed-numerologies in time in terms of vacating the queues of mMTC users. Importantly, it achieves the lowest power consumption as compared to the mixed numerologies in time and fixed numerology. Therefore, a slice-aware scheduling scheme with mixed-numerologies in frequency is the right choice to satisfy users' requirements from different services with less power consumption.

VI. CONCLUSION AND FUTURE DIRECTIONS

In this article, we have studied the slice-isolation, and slice-aware RAN radio resource assignment schemes for the efficient multiplexing of eMBB, URLLC, and mMTC services by considering the mixed-numerology based time-frequency resource grids. Mainly, we have considered the joint power and RBs allocation problems and formulated these resource allocation problems as power minimization problems to minimize the power consumption at the BS while satisfying the different QoS requirements of the users from the three services. Besides, the channel outage probability model was considered to study the impact of the imperfect CSI on the proposed radio resource scheduling algorithms. The formulated problems are combinatorial and non-convex optimization problems, which are complex to solve. In this regard, we provided low-complexity sub-optimal solutions to the formulated problems using the penalized formulation, DC programming and successive convex approximation.

We illustrated the performance of slice-isolation and slice-aware resource scheduling schemes with the different time-frequency resource grids through the simulation results. With the mixed-numerologies in the frequency domain, the slice-aware scheduling scheme achieves better performance in terms of eMBB data rates, URLLC packet latencies, and mMTC queues by consuming the lower power.

By exploiting the multiple antennas at the BS, we can easily extend our work to investigate the optimization of the number of antennas jointly with the transmit power and resource blocks of mixed-numerologies to the eMBB, URLLC, and mMTC users for the energy-efficiency maximization. Specifically, different beamforming techniques are necessary to satisfy URLLC, eMBB, and mMTC users' QoS requirements. Therefore, a study on RAN slicing problems by considering beamforming techniques and frame structures with mixed numerologies can be also envisaged to analyze its performance in the next-generation networks.

Future works must investigate the optimal data buffer sizes, TCP flow and congestion control mechanisms to avoid the buffer bloat problems in the 5G wireless networks.

REFERENCES

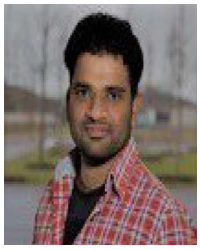
[1] J. G. Andrews *et al.*, "What will 5G be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[2] "IMT vision-framework and overall objectives of the future development of IMT for 2020 and beyond, ITU recommendation," ITU-R, Geneva, Switzerland, ITU-Recommendation M. 2083-0, Sep. 2015.

[3] *Cloud VR Bearer Networks*, Huawei iLab, Shenzhen, China, 2017. [Online]. Available: https://www-file.huawei.com/-/media/CORPORATE/PDF/ilab

[4] "Feasibility study on new services and markets technology enablers-enhanced mobile broadband," 3GPP, Sophia Antipolis, France, 3GPP Rep. TR22.863, Sep. 2016.

[5] P. Schulz *et al.*, "Latency critical IOT applications in 5G: Perspective on the design of radio interface and network architecture," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 70–78, Feb. 2017.
[6] "5G PPP white papers on energy, automotive, factories, and eHealth vertical sectors," The 5G Infrastruct. Public Private Partnership, Heidelberg, Germany, White Paper, Aug. 2016. [Online]. Available: https://5g-ppp.eu/white-papers
[7] M. Simsek, A. Aijaz, M. Dohler, J. Sachs, and G. Fettweis, "5G-enabled tactile Internet," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 3, pp. 460–473, Mar. 2016.
[8] S. K. Sharma and X. Wang, "Toward massive machine type communications in ultra-dense cellular IOT networks: Current issues and machine learning-assisted solutions," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 1, pp. 426–471, 1st Quart., 2020.
[9] "Feasibility study on new services and markets technology enablers for massive Internet of Things," 3GPP, Sophia Antipolis, France, 3GPP Rep. TR22.861, Sep. 2016.
[10] "Study on new radio (NR) access technology physical layer aspects," 3GPP, Sophia Antipolis, France, 3GPP Rep. TR38.802v14.0.0, Mar. 2017.
[11] H. Zhang, N. Liu, X. Chu, K. Long, A. Aghvami, and V. C. M. Leung, "Network slicing based 5G and future mobile networks: Mobility, resource management, and challenges," *IEEE Commun. Mag.*, vol. 55, no. 8, pp. 138–145, Aug. 2017.
[12] 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.
[13] L. Zhang, A. Ijaz, P. Xiao, A. Quddus, and R. Tafazolli, "Subband filtered multi-carrier systems for multi-service wireless communications," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1893–1907, Mar. 2017.
[14] A. Yazar and H. Arslan, "Flexible multi-numerology systems for 5G new radio," *J. Mobile Multimedia*, vol. 14, no. 2, pp. 367–394, 2018.
[15] S. Lien, S. Shieh, Y. Huang, B. Su, Y. Hsu, and H. Wei, "5G new radio: Waveform, frame structure, multiple access, and initial access," *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 64–71, Jun. 2017.
[16] B. Khodapanah, A. Awada, I. Viering, J. Francis, M. Simsek, and G. P. Fettweis, "Radio resource management in context of network slicing: What is missing in existing mechanisms?" in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Marrakesh, Morocco, 2019, pp. 1–7.
[17] G. Pocovi, K. I. Pedersen, and P. Mogensen, "Joint link adaptation and scheduling for 5G ultra-reliable low-latency communications," *IEEE Access*, vol. 6, pp. 28912–28922, 2018.
[18] 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. (VTC2019-Spring)*, Kuala Lumpur, Malaysia, 2019, pp. 1–6.
[19] P. Popovski, K. F. Trillingsgaard, O. Simeone, and G. Durisi, "5G wireless network slicing for eMBB, URLLC, and mMTC: A communication-theoretic view," *IEEE Access*, vol. 6, pp. 55765–55779, 2018.
[20] A. Anand, G. de Veciana, and S. Shakkottai, "Joint scheduling of URLLC and eMBB traffic in 5G wireless networks," *IEEE/ACM Trans. Netw.*, vol. 28, no. 2, pp. 477–490, Apr. 2020.
[21] 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.
[22] L. You, Q. Liao, N. Pappas, and D. Yuan, "Resource optimization with flexible numerology and frame structure for heterogeneous services," *IEEE Commun. Lett.*, vol. 22, no. 12, pp. 2579–2582, Dec. 2018.
[23] 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.
[24] P. Guan *et al.*, "5G field trials: OFDM-based waveforms and mixed numerologies," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 6, pp. 1234–1243, Jun. 2017.
[25] T. Bag, S. Garg, Z. Shaik, and A. Mitschele-Thiel, "Multi-numerology based resource allocation for reducing average scheduling latencies for 5G NR wireless networks," in *Proc. Eur. Confe. Netw. Commun. (EuCNC)*, 2019, pp. 597–602.
[26] C. Sexton, N. Marchetti, and L. A. DaSilva, "Customization and trade-offs in 5G RAN slicing," *IEEE Commun. Mag.*, vol. 57, no. 4, pp. 116–122, Apr. 2019.
[27] 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)*, Shanghai, China, 2019, pp. 1–6.
[28] L. Marijanovic, S. Schwarz, and M. Rupp, "Multi-user resource allocation for low latency communications based on mixed numerology," in *Proc. IEEE 90th Veh. Technol. Conf. (VTC-Fall)*, Honolulu, HI, USA, 2019, pp. 1–7.
[29] P. Korrai, E. Lagunas, S. K. Sharma, S. Chatzinotas, A. Bandi, and B. Ottersten, "A RAN resource slicing mechanism for multiplexing of eMBB and URLLC services in OFDMA based 5G wireless networks," *IEEE Access*, vol. 8, pp. 45674–45688, 2020.
[30] D. W. K. Ng, E. S. Lo, and R. Schober, "Energy-efficient resource allocation in multi-cell OFDMA systems with limited backhaul capacity," *IEEE Trans. Wireless Commun.*, vol. 11, no. 10, pp. 3618–3631, Oct. 2012.
[31] N. Le, L. Tran, Q. Vu, and D. Jayalath, "Energy-efficient resource allocation for OFDMA heterogeneous networks," *IEEE Trans. Commun.*, vol. 67, no. 10, pp. 7043–7057, Oct. 2019,
[32] "Cellular system support for ultra low complexity and low throughput Internet of Things," 3GPP, Sophia Antipolis, France, 3GPP TR 45.820v13.0.0, Aug. 2015.
[33] A. B. Kihero, M. S. J. Solaija, and H. Arslan, "Inter-numerology interference for beyond 5G," *IEEE Access*, vol. 7, pp. 146512–146523, 2019.
[34] D. W. K. Ng and R. Schober, "Cross-layer scheduling for OFDMA amplify-and-forward relay networks," *IEEE Trans. Veh. Technol.*, vol. 59, no. 3, pp. 1443–1458, Mar. 2010.
[35] S. S. Ikki and S. Aissa, "Two-way amplify-and-forward relaying with Gaussian imperfect channel estimations," *IEEE Commun. Lett.*, vol. 16, no. 7, pp. 956–959, Jul. 2012.
[36] N. Mokari, M. R. Javan, and K. Navaie, "Cross-layer resource allocation in OFDMA systems for heterogeneous traffic with imperfect CSI," *IEEE Trans. Veh. Technol.*, vol. 59, no. 2, pp. 1011–1017, Feb. 2010.
[37] D. Lopez-Perez, X. Chu, A. V. Vasilakos, and H. Claussen, "Power minimization based resource allocation for interference mitigation in OFDMA femtocell networks," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 2, pp. 333–344, Feb. 2014.
[38] J. Lee and S. Leyffer, *Mixed Integer Non-Linear Programming*. New York, NY, USA: Springer, 2011.
[39] F. Fang, H. Zhang, J. Cheng, S. Roy, and V. C. M. Leung, "Joint user scheduling and power allocation optimization for energy-efficient NOMA systems with imperfect CSI," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2874–2885, Dec. 2017.
[40] A. Bandi and C. R. Murthy, "Structured sparse recovery algorithms for data decoding in media based modulation," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2017, pp. 1–6.
[41] M. Grant and S. Boyd. (2011). *CVX: MATLAB Software for Disciplined Convex Programming*. [Online]. Available: http://cvxr.com/cvx
[42] P. Gahinet, A. Nemirovski, A. J. Laub, and M. Chilali, *LMI Control Toolbox Users Guide*. Natick, MA, USA: MathWorks, 1995.



PRAVEEN KUMAR KORRAI (Student Member, IEEE) received the M.A.Sc (T) degree from the Department of Electrical and Computer Engineering, Concordia University, Montreal, Canada. He is currently pursuing the Ph.D. degree with the Interdisciplinary Center for Security, Reliability and Trust (SnT), University of Luxembourg, Luxembourg. He has working experience as a Researcher. He holds a grant for his Ph.D. project received from Luxembourg National Research Fund (FNR), under Individual Ph.D. Fellowship Scheme. His research interests are cognitive communications, machine learning, and millimeter wave communications, performance evaluation of wireless networks, sparse signal processing techniques, and FPGA implementation of wireless communication techniques.

EVA LAGUNAS (Senior Member, IEEE) received the M.Sc. and Ph.D. degrees in telecommunications engineering from the Polytechnic University of Catalonia (UPC), Barcelona, Spain, in 2010 and 2014, respectively, where she was a Research Assistant with the Department of Signal Theory and Communications, from 2009 to 2013. During the summer of 2009, she was a Guest Research Assistant with the Department of Information Engineering, Pisa, Italy. From November 2011 to May 2012, she held a visiting research appointment with the Center for Advanced Communications, Villanova University, Villanova, PA, USA. In 2014, she joined the Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg, where she currently holds a Research Scientist position. Her research interests include radio resource management and general wireless networks optimization.

ASHOK BANDI was born in Kunkalagunta, India, in 1988. He received the M.Tech. degree in electronics and communication engineering from the National Institute of Technology (NIT), Tiruchirappalli, India, in 2012. He is currently pursuing the Ph.D. degree in electrical engineering with the University of Luxembourg, Luxembourg. He has worked on physical layer design and development for WLAN 802.11a/n/ac with Imagination Technologies, Hyderabad, India, from 2012 to 2015, and with National Instruments, Bengaluru, India, from 2015 to 2016. He was worked as a Project Associate with the Department of ECE, IISc, Bengaluru, from 2016 to May 2017. He joined the Interdisciplinary Centre for Security, Reliability, and Trust, University of Luxembourg, in June 2017. He is working on sparse signal recovery and joint update of integer and non-linear variables in MINLP problems that appear in for wireless communications within the Project PROSAT (on-board PROcessing techniques for high throughput SATellites), funded under FNR CORE-PPP Framework.

SHREE KRISHNA SHARMA (Senior Member, IEEE) received the Ph.D. degree in wireless communications from the University of Luxembourg in 2014, where he is currently a Research Scientist with the SnT. Prior to this, he worked as a Postdoctoral Fellow with the University of Western Ontario, Canada, in the areas of 5G wireless communications and Internet of Things systems. He also worked as a Research Associate with the SnT being involved in different European, national and ESA projects. In the past, he had held an industrial position as a Telecom Engineer with Nepal Telecom, and part-time and full-time teaching positions with three different universities in Nepal. He has published more than 90 technical papers in scholarly journals and international conferences, and has over 1600 Google scholar citations. His current research interests include 5G and beyond wireless, Internet of Things, machine learning, edge computing, and optimization of distributed communications, computing, and caching resources. He is the recipient of the several prestigious awards, including the 2018 EURASIP Best Journal Paper Award, the Best Paper Award in CROWNCOM 2015 Conference, and the FNR Award for Outstanding Ph.D. Thesis 2015 from the FNR, Luxembourg. He has been serving as a Reviewer for several international journals and conferences; as a TPC Member for a number of international conferences, including IEEE ICC, IEEE GLOBECOM, IEEE PIMRC, IEEE VTC, and IEEE ISWCS; and an Associate Editor for IEEE ACCESS journal. He organized a special session in IEEE PIMRC 2017 Conference, worked as a Track Co-Chair for IEEE VTC-Fall 2018 Conference, and has recently published an edited book on *Satellite Communications in the 5G Era* with the IET as a Lead Editor.

SYMEON CHATZINOTAS (Senior Member, IEEE) received the M.Eng. degree in telecommunications from the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 2003, and the M.Sc. and Ph.D. degrees in electronic engineering from the University of Surrey, Surrey, U.K., in 2006 and 2009, respectively. He is currently the Co-Head of the SIGCOM Research Group, Interdisciplinary Centre for Security, Reliability, and Trust, University of Luxembourg, Luxembourg, and also a Visiting Professor with the University of Parma, Italy. He was involved in numerous research and development projects for the Institute of Informatics Telecommunications, National Center for Scientific Research Demokritos, the Institute of Telematics and Informatics, Center of Research and Technology Hellas, and the Mobile Communications Research Group, Center of Communication Systems Research, University of Surrey. He has more than 300 publications, 3000 citations, and an H-Index of 30 according to Google Scholar. He was a co-recipient of the 2014 IEEE Distinguished Contributions to Satellite Communications Award, the CROWNCOM 2015 Best Paper Award, and the 2018 EURASIP JWCN Best Paper Award.