IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 29, 2020, accepted November 15, 2020, date of publication November 19, 2020,
date of current version December 3, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3039352

Multiplexing Services in 5G and Beyond: Optimal Resource Allocation Based on Mixed Numerology and Mini-Slots

LJILJANA MARIJANOVIĆ[1,2], (Student Member, IEEE), STEFAN SCHWARZ[1,2], (Senior Member, IEEE), AND MARKUS RUPP[1], (Fellow, IEEE)

[1]Institute of Telecommunications, Vienna University of Technology, 1041 Vienna, Austria
[2]Christian Doppler Laboratory for Dependable Wireless Connectivity for the Society in Motion, 1040 Vienna, Austria

Corresponding author: Ljiljana Marijanović (ljiljana.marijanovic@tuwien.ac.at)

This work has been funded by the Christian Doppler Laboratory for dependable Wireless Connectivity for the Society in Motion. The financial support by the Austrian Federal Ministry for Digital and Economic Affairs, the National Foundation for Research, Technology and Development and the Christian Doppler Research Association is gratefully acknowledged. The authors acknowledge the TU Wien University Library for financial support through its Open Access Funding Programme.

ABSTRACT The Fifth Generation (5G) New Radio (NR) Physical Layer (PHY) is designed to successfully address diverse user and service requirements by providing a highly flexible framework. This flexibility is viable through a scalable numerology. Since 5G NR targets to multiplex various applications with different quality of service requirements within the same band, 3rd Generation Partnership Project has introduced a *mixed (multi) numerology* approach and a *mini-slot* approach to enhance the adaptability of the PHY. In this contribution, we compare these two approaches focusing on the achievement of low-latency communications. We propose optimization problems that enable to maximize the achievable rate of best effort users, while maintaining latency requirements of low-latency users. Exploiting achievable rate performance as one of the fundamental metrics, we show a comparison of *mixed numerology* and *mini-slot* approach in different circumstances. In addition to Cyclic Prefix (CP)-Orthogonal Frequency Division Multiplexing (OFDM), we employ Universal Filtered Multicarrier (UFMC) as a potential beyond 5G technology and show that it achieves an improvement over CP-OFDM. The optimization problems for both mixed numerology and the mini-slot approach are initially given by an integer programming solution. In order to reduce computational complexity for large-scale scenarios, we apply the *Dantzig-Wolfe* decomposition method, showing that it is possible to achieve the optimal solution with significantly reduced complexity by exploiting the structure of the proposed optimization formulation.

INDEX TERMS OFDM, optimal scheduling, physical layer, resource management, scalability, 5G mobile communication.

I. INTRODUCTION

The Fifth Generation (5G) of mobile communications and beyond is expected to cope with various requirements and services, and hence its air interface has to provide a highly flexible configuration. 5G operates on a wide range of frequencies, from several hundreds of MHz to the millimeter-wave bands, and encourages new functionalities in order to support human-centric and machine-centric applications [1], [2]. To cope with a large variety in demands of different applications, 3rd Generation Partnership Project (3GPP) has identified three main use-case families: Enhanced Mobile Broadband (eMBB), Massive Machine-Type Communications (mMTC) and Ultra-Reliable Low-Latency Communications (uRLLC) [3]–[5]. Thereby applications with the same or similar demands are united into one group within the corresponding use-case family. uRLLC and mMTC require a Low-Latency (LL) as well as high reliability performance (process automation, virtual reality etc.) and high connection density of devices (smart cities, smart building applications etc.), respectively, [6]–[11]. On the other hand, eMBB supports a high data rate, increased traffic volume and higher

The associate editor coordinating the review of this manuscript and approving it for publication was Jie Tang.

user mobility (high-speed internet access, multimedia content etc.). Nevertheless, such applications with similar demands may exhibit very different channel conditions. For example, both an indoor hotspot scenario and a high-speed vehicle scenario may require very high data rate, thus belonging to the eMBB group. However, they experience dissimilar channel conditions; indoor scenarios exhibit small Doppler and delay spreads, whereas high-speed outdoor environments exhibit large Doppler and delay spreads. To satisfy different requirements imposed by eMBB, mMTC and uRLLC traffic, 5G New Radio exploits a scalable numerology. In addition to facilitating an efficient support of various service requirements, employing a scalable numerology allows a better adjustment of the Physical Layer (PHY) waveform to different channel conditions, providing more robustness against channel variations [12].

The term numerology refers to the PHY waveform parametrization [2]. Unlike Long Term Evolution (LTE), which supports a single numerology in downlink, 5G enables a scalable numerology, referring to the choice of different subcarrier spacings, symbol durations and number of symbols per time slot. The baseline 15 kHz subcarrier spacing is scalable by a factor 2^{s-1}, where s is a non-negative integer ranging from 1 to 4. The symbol duration is scaled inversely with subcarrier spacing, and hence the number of symbols that can be arranged within a fixed time slot is different among numerologies. To support multiple services on the same carrier, 3GPP proposes a *mixed numerology* for frequency multiplexing and a *mini-slot* approach that employs a single predefined numerology with shorter slot durations than a regular slot for that numerology [13], [14].

One particular focus of 5G NR is on the achievement of LL communications [5]. Both mixed numerology and mini-slot methods can be used for LL transmissions. For example, a mini-slot approach, as the name implies, enables transmissions that are shorter than the regular slot duration, supporting LL and eMBB services at the same time. On the other hand, LL can also be accomplished by employing wider subcarrier spacings, since in such a case slots have a shorter duration than an LTE slot. In this work, the main focus is on the latency aspect of uRLLC. The reliability aspect is not considered. The reliability aspect could be achieved within our proposed framework by applying a sufficiently large fading margin to the Signal-to-Interference and Noise Ratio (SINR)/ achievable rate to ensure reliable transmission. The approach to solving the problem in this case would remain the same.

Recently, significant progress has been made in the context of resource management of flexible and mixed numerology transmissions. To emphasize the importance of our work as well as the value of the contribution, we compare it with similar works on this topic. A comprehensive study on optimal resource allocation in uRLLC for real-time wireless control systems is proposed in [15]. In [16], the authors propose an efficient resource allocation scheme for maximizing the energy efficiency. However, they do not consider the mixed numerology concept and hence specifically no Internumerology Interference (INI) and no partitioning into subbands. In [17], the authors provide a heuristic algorithm to decide on the efficient number of concurrently active mixed numerologies. The authors choose a trade-off between system requirements and channel conditions by introducing so-called necessity weights. For example, for a *highway* scenario, the necessity for a large subcarrier spacing is high and therefore they select it. However, this approach is not proven to be optimal. In contrast, we propose an accurate SINR model that accounts for all relevant PHY characteristics of a mixed numerology system, and utilize it to optimize the numerology and resource allocation under latency constraints. In [18] and [19] the authors consider an optimal resource assignment under different channel conditions within a mixed numerology approach. The authors in [20] investigate the resource allocation taking into account Intersymbol Interference (ISI) and Intercarrier Interference (ICI), but in contrast to our work, INI and the channel estimation error are not considered. INI is the main drawback of the mixed numerology approach and it certainly affects the optimizer decision and therefore it has to be considered. A flexible framework for a multi-service system based on subband filtered multicarrier transmission is established in [21]. The authors reduce interference between numerologies by applying cancellation and equalization algorithms. According to [22], frequency multiplexing of uRLLC and eMBB services within one carrier is feasible by employing subband-filtered Cyclic Prefix (CP)- Orthogonal Frequency Division Multiplexing (OFDM). By setting different objectives, the authors in [23]–[25] investigate efficient resource scheduling algorithms for supporting heterogeneous services based on a flexible frame structure. The work presented in [23] considers the scheduling problem for heterogeneous services within a mixed numerology approach. It aims to maximize the number of satisfied users while meeting latency demand and data transmission requirements. In contrast to our work, the authors in [23] do not consider a joint optimization of the numerology and resource allocation; they rather predefine the numerology of each user based on the channel conditions, but this approach is heuristic. Of course, in such cases suboptimal numerology degrades performance. The authors in [25] present QoS-aware resource allocation utilizing the flexibility in terms of mixed numerology. Similarly to many works on this topic and contrary to our work, their solution is based on heuristic approaches that do not accurately consider PHY characteristics. In [26] and [27] the authors evaluate the shortcomings and advantages of the mini-slot concept. By applying machine learning-based algorithms, the authors in [26] propose a flexible Transmission Time Interval (TTI)[1] scheduling in eMBB and uRLLC coexistence scenarios. TTI-based scheduling is also investigated in [27] and the optimal selection of TTI is studied at different system loads. Since mixed numerology and the mini-slot approach both target very similar use cases, it is essential to investigate

[1] In this paper, we use the term *slot* which is equivalent to the term TTI.

the merits and drawbacks of both methods in detail. In [28], system-level simulations are conducted in order to evaluate the performance of the mixed-numerology concept in conjunction with mini-slots. Based solely on latency requirements, the authors propose a dynamic resource allocation scheme to reduce the average scheduling latency. The authors in [29] elaborate the optimal resource allocation for the mixed numerology approach.

A. CONTRIBUTIONS

In this paper, we investigate an optimal numerology and resource allocation for mixed numerology and the mini-slot approach. Although a combination of these two methods has been investigated by some researchers, there is no research that cross-compares the advantages and drawbacks of both of these methods separately. Our main goal is to investigate which of these approaches is better suited for LL communications, the mixed numerology approach with its drawbacks such as INI, or the mini-slot approach with shorter slot durations but fixed numerology. The main difference between our work and relevant works on this topic is that we base our optimization not only on QoS requirements (throughput and LL), but we also include PHY characteristics (ICI, ISI, channel estimation error and INI), which characterize the channel conditions and transceiver performance. Therefore, our obtained solutions are the results of a combination of user/service requirements and channel conditions. The main contributions are:

- We investigate the optimal time-frequency resource allocation in a multi-user scenario employing mixed numerology and mini-slot methods. We propose novel optimization methods for joint numerology and resource allocation in mixed numerology systems, as well as, for numerology selection and mini-slot allocation in mini-slot systems. Unlike many related papers on this topic that exclusively consider QoS requirements, such as LL or high throughput imposed by the application, we present a more realistic scenario, by additionally taking into account the impact of PHY parameters, including interference induced by the channel, channel estimation errors and INI. These PHY characteristics can have an even larger impact on the resource and numerology allocation than the QoS requirements. All the necessary constraints concerning the packet size and latency are analyzed in detail. We provide a comprehensive comparison of the mixed numerology and mini-slot approaches under various latency constraints and channel conditions, underlining the preferable option in all investigated cases.
- Besides the scalable frame structure proposed for 5G, the flexibility for beyond 5G technologies can potentially be increased by employing different waveform contenders. In addition to CP-OFDM proposed in 5G NR downlink, we also consider Universal Filtered Multicarrier (UFMC) as a filtered version of OFDM for beyond 5G and investigate the impact of this waveform on the optimization formulation. We analyze the effect of better spectral confinement on interference between numerologies. Thereby, we evaluate the impact of the waveform on the numerology selection as well as the user resource allocation. Moreover, we compare it to CP-OFDM in order to comprehend the potential gain.
- Our optimization problems for both methods are formulated as multiscenario max-min *knapsack* problems which can be solved by integer linear programming (ILP). In addition to the ILP solver that provides optimal solutions of the proposed optimization problems with a high computational complexity for large-scale scenarios, we show that the *Dantzig-Wolfe* decomposition can be applied to our optimization problem in order to reduce complexity. This algorithm substantially reduces the complexity, by performing the computations on subblocks of the original problem rather than on the full size problem.

B. ORGANIZATION

In Sec. II-A and Sec. II-B, we present our system model employing both, CP-OFDM and UFMC, respectively. We analyze the system SINR including all interference caused either by the channel variations or the non-orthogonality between different numerologies. Furthermore, we model the estimation error under the assumption that a least-squares channel estimator is used. In Sec. II-C, we provide a brief review of the numerology as well as the Resource Block (RB) definition proposed by 3GPP. In Sec. III, we formulate the optimization problems for the mixed numerology and the mini-slot approach and describe in detail the corresponding constraints required to satisfy all user and service demands. In order to reduce the computational complexity of our optimization, we apply the *Dantzig-Wolfe* decomposition method in Sec. IV. We consolidate and explain all results in Sec. V. Finally, we conclude our work in Sec. VI.

C. NOTATIONS

We denote a complex Gaussian distribution with zero mean and variance σ^2 with $\mathcal{CN}(0, \sigma^2)$. With $|x|$ we denote the absolute value of complex-valued scalar x, with $|\mathcal{X}|$ we denote the size of cardinality $\mathcal{X}$ and with $\mathbb{X}$ the mathematical expectation of random variable X.

II. SYSTEM MODEL

In Sec. II-A and Sec. II-B, we provide an input-output relationship of the single-user transmission employing two waveforms, CP-OFDM and UFMC, respectively. We provide a detailed analysis of SINR including interference caused by the channel variations, i.e., ICI and ISI, as well as, the interference caused by mixed numerology, i.e., INI. Furthermore, we assume imperfect channel knowledge and hence we account for the channel estimation error. We also consider Additive White Gaussian Noise (AWGN). In addition to the SINR analysis, in Sec. II-C we provide a brief review of

the scalable numerology proposed by 3GPP and clarify the time-frequency resource management.

A. CP-OFDM

The input-output relationship of a single-user CP-OFDM transmission is given by:[2]

$$Y_{k,n}^{\text{ofdm}} = H_{k,n} X_{k,n}^{\text{ofdm}} + W_{k,n}^{\text{ofdm}}, \tag{1}$$

where $X_{k,n}^{\text{ofdm}}$ is the transmitted data symbol, $H_{k,n}$ is the frequency response of the system and $Y_{k,n}^{\text{ofdm}}$ denotes the received data symbol at frequency-time index k, n. ICI power ($\sigma^{2,\text{ofdm}}_{\text{ICI}_{\Delta f}}$), ISI power ($\sigma^{2,\text{ofdm}}_{\text{ISI}_{\Delta f}}$), INI power ($\sigma^{2,\text{ofdm}}_{\text{INI}_{\Delta f,k}}$), the channel estimation error ($\sigma^{2,\text{ofdm}}_{\text{e}_{\Delta f}}$), as well as the noise power ($\sigma^{2,\text{ofdm}}_{\text{n}}$) can approximately be modeled as $W_{k,n}^{\text{ofdm}} \sim \mathcal{CN}(0, \sigma^{2,\text{ofdm}}_{\text{ICI}_{\Delta f}} + \sigma^{2,\text{ofdm}}_{\text{ISI}_{\Delta f}} + \sigma^{2,\text{ofdm}}_{\text{INI}_{\Delta f,k}} + \sigma^{2,\text{ofdm}}_{\text{e}_{\Delta f}} + \sigma^{2,\text{ofdm}}_{\text{n}})$ [30]. The input-output relationship is normalized by the pathloss, such that different users may have different noise variance.

We consider a diverse multipath environment, as well as, movement of users. Correspondingly, the channels of different users are characterized by different Root-Mean-Square (RMS) delay and Doppler spreads. These channel characteristics can cause ISI and ICI, respectively. Specifically, due to large Doppler spreads compared to the subcarrier spacing (Δf), the system faces ICI, and due to insufficient CP lengths $T_{\text{CP},\Delta f}$ compared to the channel delay spread, the system is subject to ISI. Both ICI and ISI are subcarrier spacing dependent metrics. ICI depends on the Doppler spread and symbol duration, whereas ISI depends on the power delay profile of the transmission channel and CP length, [31], [32]. Intuitively, larger subcarrier spacings are more robust to ICI than smaller subcarrier spacings, but less resistant to ISI due to shorter CP lengths. We assume that each of the users may experience different channel conditions. That is, we consider a mix of indoor/outdoor and low/high mobility users, which are therefore subject to different delay and Doppler spreads of the wireless channel. The delay and Doppler spread of each user determine the powers of ISI $\sigma^2_{\text{ISI}_{\Delta f}}$ and ICI $\sigma^2_{\text{ICI}_{\Delta f}}$, respectively.

We model the channel estimation error $\sigma^2_{\text{e}_{\Delta f}}$ as proposed in [33] under the assumption that a least-squares channel estimator with a 2D linear interpolation is utilized. However, we are not restricted only to this channel estimator. We can use any channel estimator, as long as the channel estimation error $\sigma^2_{\text{e}_{\Delta f}}$ can be characterized. As shown in [33], the channel estimation error $\sigma^2_{\text{e}_{\Delta f}}$ can be stated in terms of the overall interference caused by the channel, as well as, INI due to the non-orthogonal system, and noise variance.

Mixed numerology enables a high flexibility, but has also some drawbacks. One of the main disadvantages is INI. Placing two users that employ different subcarrier spacings (numerologies) next to each other leads to leakage of the spectrum of one user onto the spectrum of the other user. As a result, we have non-orthogonal subcarriers between the users causing interference [34]. We assume a multi-user system with time-frequency sharing of resource elements amongst users. Hence, users that occupy the same numerology are orthogonal, whereas, transmissions with different numerologies are subject to INI. We use the term *victim* user and the term *aggressor* user to denote the user affected by INI and the user that causes INI, respectively. For more details about ICI and ISI, channel estimation error and INI, we refer the reader to Appendix A.

In Fig. 1, we show an example of INI power caused by an *aggressor* user towards a *victim* user. The INI power is represented in terms of the spectrum of the *aggressor* user within the bandwidth of the *victim* user. In this specific example, the *aggressor* user employs either 30 kHz, 60 kHz or 120 kHz subcarrier spacing and the *victim* user employs 15 kHz subcarrier spacing. From Fig. 1, we can observe that the interference is zero whenever the 15 kHz subcarriers fall onto a multiple of 30 kHz, 60 kHz and 120 kHz, respectively.

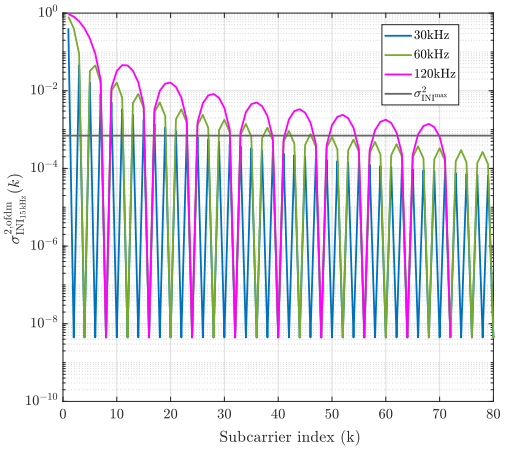


FIGURE 1. INI caused by 30 kHz, 60 kHz and 120 kHz onto 15 kHz subcarrier spacing.

One of the challenges regarding multi-numerology systems is to find an adaptive guard band between numerologies. One way to do this is to define a spectral mask, i.e., to meet Out-Of-Band emission (OOB) requirements. The Adjacent Channel Leakage power Ratio (ACLR) and Operating Band Unwanted Emissions (OBUE) methods are most commonly used in the literature [35]. In contrast to these methods which propose more complex mask structures, we consider a simplified form of guard band definition, which simplifies the solution of the optimization problems formulated in Sec. III. We define a spectral mask as an upper bound on allowed INI, $\sigma^2_{\text{INI}^{\max}}$, as shown in Fig. 1. By defining the spectral mask we are able to define a size of the guard band between numerologies, i.e., the number of subcarriers required to separate numerologies in order to reduce INI on the victim user below this threshold. The selection of this threshold value is a tradeoff between guard band overhead and INI. That is, smaller $\sigma^2_{\text{INI}^{\max}}$ implies a larger guard band overhead and therefore reduced spectral efficiency. Yet, with more

[2]Here we omit a user index for brevity and introduce it in Sec. III when we consider multi-user resource allocation.

guard subcarriers we have less severe INI. We select this upper bound such that we prevent a dominant effect of INI compared to the other interference and noise sources. In other words, with the guard band defined by this threshold we mitigate the strongest INI, but still keep the spectral efficiency high.

Finally, including the aforementioned parameters we can write the SINR expression as[3]:

$$\bar{\gamma}_{\Delta f} \geq \frac{\sigma_{\mathrm{d}}^2 \sigma_H^2}{\sigma_{\mathrm{n}}^2 + \sigma_{\mathrm{ICI}_{\Delta f}}^2 + \sigma_{\mathrm{ISI}_{\Delta f}}^2 + \sigma_{\mathrm{e}_{\Delta f}}^2 + \sigma_{\mathrm{INI}_{\Delta f}^{\max}}^2}, \tag{2}$$

where σ_{d}^2 denotes the power of data symbols and σ_H^2 denotes the channel power. We consider the long-term average SINR, which only depends on channel statistics σ_H^2, avoiding the dependency of the numerology and resource allocation on the instantaneous channel state. As mentioned above, we consider a normalized input-output relationship (w.r.t. the pathloss); hence $\sigma_H^2 = 1$. Considering the average SINR is reasonable for numerology allocations, since it is practically not feasible to constantly vary the numerology from one slot (TTI) to the next.

The mini-slot approach is not impaired by INI; hence, its SINR is:

$$\bar{\gamma}_{\Delta f} = \frac{\sigma_{\mathrm{d}}^2}{\sigma_{\mathrm{n}}^2 + \sigma_{\mathrm{ICI}_{\Delta f}}^2 + \sigma_{\mathrm{ISI}_{\Delta f}}^2 + \sigma_{\mathrm{e}_{\Delta f}}^2}. \tag{3}$$

This implies that the SINR of the mini-slot approach is slightly enhanced compared to the mixed numerology.

B. UFMC

UFMC performs filtering at the transmitter side on small groups of consecutive subcarriers (subbands). The subband-wise filtering is intended to improve spectral characteristics and therefore to reduce INI between adjacent subbands. There are many transceiver structures of UFMC proposed in the literature. Some of them employ either Zero Padding (ZP) or CP, whereas others do not employ any guard interval between consecutive symbols in order to increase the spectral efficiency. We employ ZP, following the structure proposed in [36], [37].

We perform a $2N$-FFT at the receiver side by padding $N - L + 1$ samples at the end of the received signal, where L is the filter length and N is the FFT size. In order to obtain useful data symbols out of $2N$ samples, our received signal is downsampled by a factor of two, such that we take only even samples into account. The input-output relationship of a single user UFMC transmission is given by [38]:

$$Y_{k,n}^{\mathrm{ufmc}} = H_{k,n} F_{k,n}^{s} X_{k,n}^{s} + W_{k,n}^{\mathrm{ufmc}}, \tag{4}$$

where $F_{k,n}^{s}$ denotes the downsampled frequency response of the time domain subband filter of length L applied on the subband s. The downsampled channel frequency response and data symbols are denoted by $H_{k,n}$ and $X_{k,n}^{s}$, respectively. ISI and ICI as the channel properties, AWGN $\sim \mathcal{CN}(0, \sigma_{\mathrm{n}}^{2,\mathrm{ufmc}})$ with slightly increased noise variance compared to the time domain white noise, INI $\sigma_{\mathrm{INI}_{\Delta f,k}}^{2,\mathrm{ufmc}}$, and channel estimation error $\sigma_{\mathrm{e}_{\Delta f}}^{2,\mathrm{ufmc}}$ are comprised in $W_{k,n}^{\mathrm{ufmc}}$.

We employ ZP between consecutive symbols in UFMC in order to avoid excessive ISI in case of a time dispersive channel. As long as the combined length of the channel L_{CH} and filter length L is less than the ZP length, ISI is mitigated. Otherwise, we model ISI and ICI according to [39] by $\sigma_{\mathrm{ISI}_{\Delta f}}^{2,\mathrm{ufmc}}$ and $\sigma_{\mathrm{ICI}_{\Delta f}}^{2,\mathrm{ufmc}}$, respectively.

There are several criteria for an optimal filter design in UFMC [40]. In our implementation we employ a Dolph–Chebyshev filter since it maximizes the side-lobe attenuation [41]. A longer filter length implies better spectral confinement. However, there is a tradeoff between the spectral confinement and intrinsic interference caused by the filter with greater length than the one of the ZP. We assume that the filter length is the same length as the ZP such that intrinsic interference does not exist.

As we already emphasized, UFMC has better spectral properties than CP-OFDM due to the filtering at the transmitter side. In Fig. 2 we compare $\sigma_{\mathrm{INI}_{15\,\mathrm{kHz}}}^{2,\mathrm{ofdm}}$ and $\sigma_{\mathrm{INI}_{15\,\mathrm{kHz}}}^{2,\mathrm{ufmc}}$ with respect to the impact of 30 kHz subcarrier spacing onto 15 kHz subcarrier spacing. As we can observe, using waveforms with better spectrum confinement implies that the side lobe suppression is higher. Although INI between two users with different numerologies still exists, its amount is less when compared to CP-OFDM. This further impacts the number of guard subcarriers between numerologies and therefore the resource allocation. In [34] we derive a closed-form expression of $\sigma_{\mathrm{INI}_{\Delta f}}^{2,\mathrm{ufmc}}$, again with respect to a victim user.

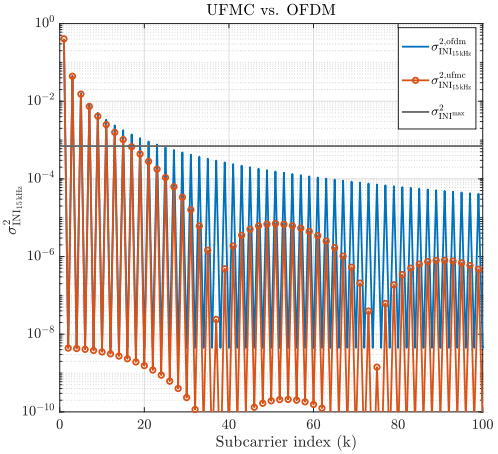
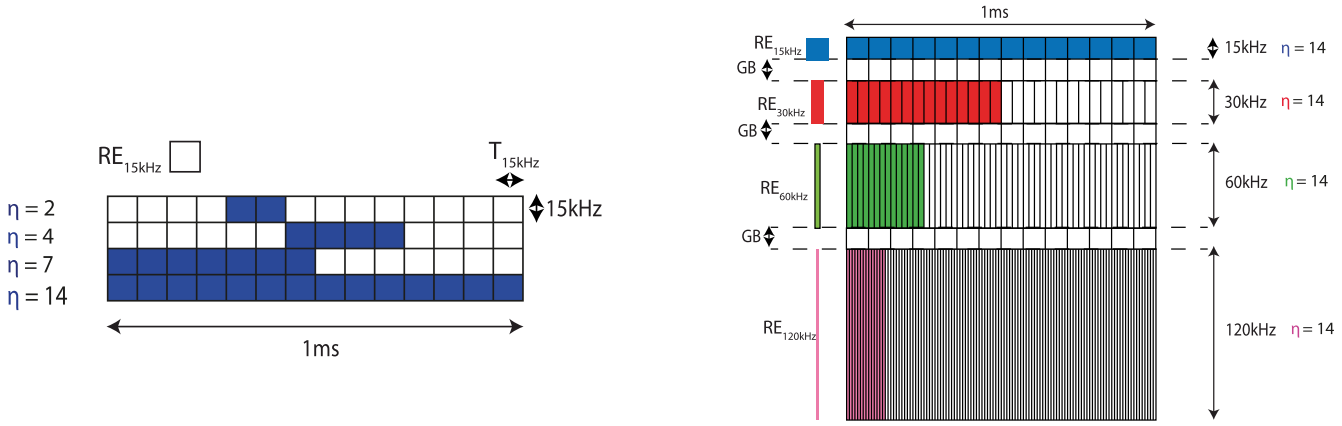


FIGURE 2. Comparison of UFMC and CP-OFDM waveforms with respect to INI.

C. TIME-FREQUENCY RESOURCE ALLOCATION

In this paper we compare two methods in order to achieve a highly flexible framework: mini-slot and mixed numerology.

[3]To simplify notations we omit the superscripts for the waveform notations.



(a) Mini-slots can have a length of 2, 4, 7 or 14 symbols.

(b) The slot duration (14 symbols) depends on Δf.

FIGURE 3. An example of mini-slot structure (a) and mixed numerology structure (b).

We consider a scalable numerology proposed by 3GPP for the bands below 6 GHz, and in the sequel we provide an overview of the time-frequency resource management for both methods.

The scalable numerology for 5G NR proposed by 3GPP is given in Table 1 [42]. The subcarrier spacings are scalable by a factor 2^{s-1} with respect to the baseline 15 kHz. The symbol duration changes inversely to the subcarrier spacing. In order to keep the same overhead as in the LTE case, that is 7 % of the total symbol duration, we also scale the length of CP/ZP. Similarly to LTE, the RB in 5G NR is composed of 12 adjacent subcarriers in frequency and 14 adjacent symbols in time (one slot). Different symbol durations imply different slot durations, i.e., smaller subcarrier spacing means longer slot duration and vice verse. However, regardless of the chosen numerology, the subframe and frame duration have fixed values of, 1 ms and 10 ms, respectively. Accordingly, the number of slots per subframe/frame changes as shown in Table 1.

TABLE 1. NR flexible numerology.

s	$s=1$	$s=2$	$s=3$	$s=4$
Subcarrier spacing [kHz]	15	30	60	120
Symbol duration [μs]	66.67	33.33	16.67	8.33
CP duration[μs]	4.69	2.34	1.17	0.59
RB frequency span [kHz]	180	360	720	1440
RB time span (slot) [ms]	1	0.5	0.25	0.125
Number of slots per subframe	1	2	4	8
Number of slots per frame	10	20	40	80

The resource allocation for the mini-slot method is based on choosing either a single or multiple blocks to a user, where each of the blocks is composed of a single subcarrier of a certain subcarrier spacing in the frequency domain and a mini-slot duration in the time domain, such as to satisfy their QoS requirements. According to 3GPP, a mini-slot can have a duration of 2, 4 or 7 consecutive symbols in the time domain. Assuming a single subcarrier of a certain spacing, it further implies 2, 4 or 7 Resource Elements (REs), respectively. Hereafter, we use the term *weights* for these number of symbols and denote it by η. Although the mini-slot durations change for different numerologies due to a different symbol duration, η always takes the same values. Regarding the frequency domain allocation, the number of allocated subcarriers is chosen to satisfy the service requirements of the users.[4]

In Fig. 3a we show an example of the mini-slot structure for a 15 kHz subcarrier spacing. We can observe that mini-slots can be positioned asynchronously with respect to the beginning of a regular slot. However, the boundaries of mini-slots have to be aligned with the symbol boundaries.

Unlike the mini-slot case, the mixed numerology approach implies a regular slot duration, i.e., 14 consecutive symbols in the time domain. In Fig. 3b we show an example of resource allocation for all subcarrier spacings from set $\mathcal{S} = \{15\,\text{kHz}, 30\,\text{kHz}, 60\,\text{kHz}, 120\,\text{kHz}\}$. We define a basic block as consisting of the REs of one single subcarrier spacing over a regular slot duration. We observe that each of these blocks consists of 14 symbols, i.e., 14 REs. Our optimization is, therefore, based on assigning a single or multiple such blocks to each user, depending on service requirements. Notice that the corresponding weight $\eta = 14$ is the same for all subcarrier spacings. Yet, the time-frequency span of a block is different for different subcarrier spacings. Different numerologies are separated by the guard band (GB) represented here by white color.

Summarizing this section, we emphasize the main features of the two considered methods:

- A mini-slot approach requires a single subcarrier spacing for all users. The base station assigns a single or multiple blocks to a user. A block contains a single subcarrier in the frequency domain and a mini-slot in

[4]This number can be restricted to be a multiple of 12, in order to allocate entire RBs to users.

the time domain. The solution depends on the selected subcarrier spacing. In case when multiple LLUs arrive with the same latency demand, they will share the bandwidth (subcarriers) within that mini-slot.

- In a mixed numerology approach different users can utilize different subcarrier spacings, but the slot duration is fixed per subcarrier spacing. The goal of our optimization is to assign multiple blocks to each user within an appropriate numerology such that we satisfy their QoS requirements.

In general, it would be possible to combine mini-slots and mixed numerology, However, in this paper, our goal is to compare the two approaches, as they have been proposed as competing concepts by different authors before. Nevertheless, it is straightforward to generalize our methods to support mini-slots in mixed numerology systems.

III. OPTIMIZATION METHODS FORMULATION

In this part we formulate and explain the constrained optimization for the user resource allocation within appropriate numerology for both, mini-slot and mixed-numerology methods. For concreteness, we describe our optimization problems in terms of the numerology parameters specified in Table 1. However, the techniques can also be applied to any other set of numerology parameters.

Similarly to the optimization criteria proposed in [16], we employ the user achievable rate per resource element as the objective function for our optimization. It is defined as:

$$\mathrm{R}^{u}_{\Delta f} = \log_2(1 + \bar{\gamma}^{u}_{\Delta f}), \quad u \in \mathcal{U}, \tag{5}$$

where u denotes a user from set $\mathcal{U}$ of all users. In addition to $\mathrm{R}^{u}_{\Delta f}$, we furthermore constrain our optimization with respect to the latency demand τ. In this regard, we distinguish two groups of users: priority LL users (LLUs) that impose a latency constraint for the transmission of packets of size q, and non LL users (NLLUs) that do not impose a latency constraint. A single LLU is denoted by u_{LL} from set $\mathcal{U}_{\mathrm{LL}}$ of all LLUs and a single NLLU is denoted by u_{NLL} from set $\mathcal{U}_{\mathrm{NLL}}$ of all NLLUs. Set $\mathcal{U}$ in (5) comprises both groups of users, i.e., $\mathcal{U} = \mathcal{U}_{\mathrm{LL}} \cup \mathcal{U}_{\mathrm{NLL}}$. We denote the total number of LLUs by $N_{\mathrm{LL}} = |\mathcal{U}_{\mathrm{LL}}|$, and the total number of NLLUs by $N_{\mathrm{NLL}} = |\mathcal{U}_{\mathrm{NLL}}|$. The total number of users is denoted by N_{users}. For simplicity, we assume that the latency constraint τ takes values equal to the slot duration, $\tau \in \{1, 0.5, 0.25, 0.125\}$ ms that corresponds to the subcarrier spacing from set $\mathcal{S}$, respectively. We consider the same latency constraint τ for all LLUs in order to simplify the notations; however, this can easily be generalized to user-specific latency constraints. We consider the same packet size q for all LLUs.

The goal of our optimization is to investigate the optimal numerology and to allocate the optimal number of resources to users with respect to their SINR values and the latency demand τ. The optimization problems for both mixed numerology and mini-slot are formulated as multi-scenario max-min *knapsack* problems, which allows us to find global optimum solutions by employing an ILP algorithm. To simplify understanding of the optimization problems formulated below, we start with a simple example as shown in Fig. 4. In this figure, we can observe that the latency constraint predefines the number of resources that can be assigned to LLUs. Namely, as illustrated in Fig. 4, LLUs must be scheduled on resource elements that are received prior to the latency constraint. Moreover, the latency constraint impacts the number of resources of NLLUs and therefore their achievable rate. In our example in Fig. 4, we consider a latency constraint of 0.5 ms; the resources predefined for LLUs are represented by red, green and dark gray colors. However, if the total number of resources occupied by LLUs does not exhaust the total number of available resources predefined by the latency constraint τ, the remaining available resources can be assigned to NLLUs. This specific case represents the mixed numerology approach with two numerologies (30 kHz and 60 kHz) and a guard band represented by white color. For better understanding of how the resources are allocated, the term *subband* is introduced. This term has been identified with the term *numerology*, i.e., a single subband implies a single numerology, regardless of the number or the category of users. We can pool all users that use a given numerology in a contiguous subband since our optimization is based on the average rather than the instantaneous SINR, and therefore it is irrelevant for a user which part of the bandwidth is allocated to it. This simplifies the optimization and it minimizes the guard overhead, as we do not need multiple guard bands between different numerologies. This can also be applied to the mini-slot approach omitting the guard band. We can notice that a more stringent τ reduces the set of possible subcarrier spacings within a mixed numerology approach that can satisfy the latency constraint, i.e., in this specific case we cannot select the 15 kHz subcarrier spacing since its slot duration is 1 ms while the latency constraint is 0.5 ms. Similarly, for a mini-slot approach, this implies a reduced set of mini-slot lengths that can be employed to satisfy the latency constraint.

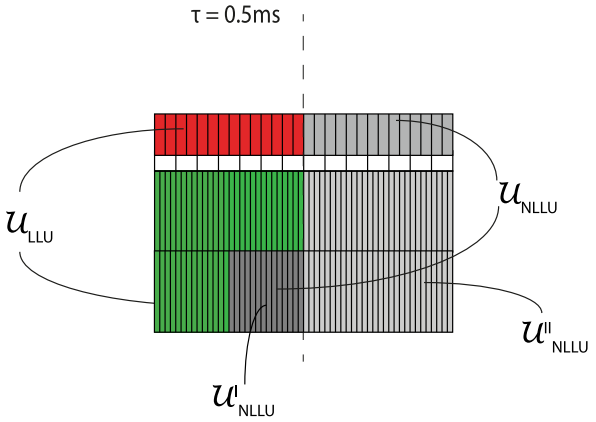


FIGURE 4. An example of the shared resources among LLUs (red and green) and NLLUs (gray), with respect to $\tau = 0.5$ ms.

A. MINI-SLOT METHOD

1) PROBLEM FORMULATION

The main goal of our optimization is to maximize the minimum achievable rate among NLLUs, ensuring at the same time that the achievable rate of LLUs is large enough to transmit packets of size q within the latency τ. The optimization problem is composed of three parts: the objective function defined in (6a), optimization variables defined in (6b) and constraints defined in (6c). It is given by:

$$\mathbf{P1}: \max \min_{u_{\mathrm{NLL}} \in \mathcal{U}_{\mathrm{NLL}}} \sum_{\eta \in \mathcal{W}} \mathrm{R}_{\Delta f}^{u_{\mathrm{NLL}}} N_{\Delta f}^{u_{\mathrm{NLL}},\eta} \eta \tag{6a}$$

$$\text{w.r.t.} \Delta f \in \mathcal{S}, N_{\Delta f}^{u_{\mathrm{NLL}},\eta} \in \mathbb{N}_0, N_{\Delta f}^{u_{\mathrm{LL}},\eta} \in \mathbb{N}_0, \quad a_{\Delta f}^{u_{\mathrm{NLL}},\eta} \in \{0,1\}, a_{\Delta f}^{u_{\mathrm{LL}},\eta} \in \{0,1\} \tag{6b}$$

$$\begin{aligned}
\text{subject to } & \sum_{\eta \in \mathcal{W}} \mathrm{R}_{\Delta f}^{u_{\mathrm{NLL}}} N_{\Delta f}^{u_{\mathrm{NLL}},\eta} \eta \geq z, \forall u_{\mathrm{NLL}} \in \mathcal{U}_{\mathrm{NLL}}, \\
& \sum_{\eta \in \mathcal{W}} \mathrm{R}_{\Delta f}^{u_{\mathrm{LL}}} N_{\Delta f}^{u_{\mathrm{LL}},\eta} \eta \geq q, \forall u_{\mathrm{LL}} \in \mathcal{U}_{\mathrm{LL}}, \\
& \sum_{\eta \in \mathcal{W}} a_{\Delta f}^{u_{\mathrm{LL}},\eta} = 1, \\
& a_{\Delta f}^{u_{\mathrm{LL}},\eta} \leq N_{\Delta f}^{u_{\mathrm{LL}},\eta} \leq a_{\Delta f}^{u_{\mathrm{LL}},\eta} N_{\max}^{\eta}, \eta \in \mathcal{W}, \\
& a_{\Delta f}^{u_{\mathrm{NLL}},\eta} \leq N_{\Delta f}^{u_{\mathrm{NLL}},\eta} \leq a_{\Delta f}^{u_{\mathrm{NLL}},\eta} N_{\max}^{\eta}, \eta \in \mathcal{W}, \\
& \sum_{u_{\mathrm{LL}} \in \mathcal{U}_{\mathrm{LL}}} \sum_{\eta \in \mathcal{W}} N_{\Delta f}^{u_{\mathrm{LL}},\eta} \eta \\
& + \sum_{u_{\mathrm{NLL}} \in \mathcal{U}_{\mathrm{NLL}}} \sum_{\eta \in \mathcal{W}} N_{\Delta f}^{u_{\mathrm{NLL}},\eta} \eta \leq C_{\mathrm{total}}, \\
& \sum_{u_{\mathrm{LL}} \in \mathcal{U}_{\mathrm{LL}}} \sum_{\eta \in \mathcal{W}} N_{\Delta f}^{u_{\mathrm{LL}},\eta} \eta \\
& + \sum_{u_{\mathrm{NLL}}^{I} \in \mathcal{U}_{\mathrm{NLL}}^{I}} \sum_{\eta \in \mathcal{W}} N_{\Delta f}^{u_{\mathrm{NLL}}^{I},\eta} \eta = C_{\mathrm{LL}}, \\
& \sum_{u_{\mathrm{NLL}}^{II} \in \mathcal{U}_{\mathrm{NLL}}^{II}} N_{\Delta f}^{u_{\mathrm{NLL}}^{II},\eta} \eta = C_{\mathrm{total}} - C_{\mathrm{LL}}, \\
& \sum_{u_{\mathrm{NLL}}^{I} \in \mathcal{U}_{\mathrm{NLL}}^{I}} N_{\Delta f}^{u_{\mathrm{NLL}}^{I},\eta} \eta + \sum_{u_{\mathrm{NLL}}^{II} \in \mathcal{U}_{\mathrm{NLL}}^{II}} N_{\Delta f}^{u_{\mathrm{NLL}}^{II},\eta} \eta, \\
& = \sum_{u_{\mathrm{NLL}} \in \mathcal{U}_{\mathrm{NLL}}} N_{\Delta f}^{u_{\mathrm{NLL}},\eta} \eta,
\end{aligned} \tag{6c}$$

with η being the mini-slot weight from set $\mathcal{W} = \{2, 4, 7, 14\}$ as shown in Fig. 3a. The total number of mini-slots with weight η occupied by a single LLU within the latency constraint τ and for subcarrier spacing Δf is denoted by $N_{\Delta f}^{u_{\mathrm{LL}},\eta}$. Similarly, the total number of mini-slots occupied by a single NLLU is denoted by $N_{\Delta f}^{u_{\mathrm{NLL}},\eta}$. Since the minimum time-resource unit is a RE, the number of mini-slots is multiplied with corresponding weight in order to obtain the total number of REs. In addition to η, $N_{\Delta f}^{u_{\mathrm{LL}},\eta}$ and $N_{\Delta f}^{u_{\mathrm{NLL}},\eta}$, the optimal subcarrier spacing Δf from set $\mathcal{S}$ is investigated. In other words, the optimization is solved for each subcarrier spacing separately, and then the most suitable one is selected in terms of the achievable rate. The binary optimization variables that indicate whether a specific NLLU/LLU transmits packets with weight η or not are denoted by $a_{\Delta f}^{u_{\mathrm{NLL}},\eta}$ and $a_{\Delta f}^{u_{\mathrm{LL}},\eta}$, respectively. The total number of available resources is denoted by C_{total} and the total number of resources predefined for LLUs is denoted by C_{LL}. They are defined as:

$$\begin{aligned}
C_{\mathrm{total}} &= K_{\Delta f} M_{\Delta f}^{1\,ms}, \\
C_{\mathrm{LL}} &= K_{\Delta f} M_{\Delta f}^{\tau},
\end{aligned} \tag{7}$$

where $M_{\Delta f}^{\tau}$ is the number of symbols of duration $T_{\Delta f}$ that can fit in latency τ, $M_{\Delta f}^{1\,ms}$ is the number of symbols that correspond to 1 ms and $K_{\Delta f}$ is the total number of subcarriers as determined by the bandwidth. The total capacity C_{total} therefore represents the total number of resource elements within a subframe. It is subcarrier spacing independent since the symbol duration $T_{\Delta f}$ and subcarrier spacing Δf are inversely proportional, such that this dependency cancels out in the multiplication in (7) [42]. In addition to the total capacity that comprises both groups of users, C_{LL} is a limit for LLUs determined by the latency constraint τ, i.e., C_{LL} represents the total number of REs available within the duration τ. We distinguish two groups of NLLUs: the first group $\mathcal{U}_{\mathrm{NLL}}^{I}$, which is interleaved with LLUs and hence shares the resources within C_{LL}, and the second group $\mathcal{U}_{\mathrm{NLL}}^{II}$ with the total number of mini-slots per user $N_{\Delta f}^{u_{\mathrm{NLL}}^{II},\eta}$ which occupies the remaining resources. The total number of mini-slots per user from the first group of NLLUs is denoted by $N_{\Delta f}^{u_{\mathrm{NLL}}^{I},\eta}$ and the total number of mini-slots per user from the second group of NLLUs is denoted by $N_{\Delta f}^{u_{\mathrm{NLL}}^{II},\eta}$.

The problem defined in (6) can be reformulated as a pure maximization problem, by introducing an auxiliary variable z, which is a lower bound on the achievable rate of NLLUs. The optimization variables as well as the constraints are already defined in (6b) and (6c), respectively. These goals are written as:

$$\begin{aligned}
\mathbf{P2}: \ & \max z \\
& \text{w.r.t.(6b)} \\
& \text{subject to } \sum_{\eta \in \mathcal{W}} \mathrm{R}_{\Delta f}^{u_{\mathrm{NLL}}} N_{\Delta f}^{u_{\mathrm{NLL}},\eta} \eta \geq z, \forall u_{\mathrm{NLL}} \in \mathcal{U}_{\mathrm{NLL}} \\
& \quad \text{(6c)}.
\end{aligned} \tag{8}$$

In the following, we discuss the constraints of this optimization problem in more detail.

2) LATENCY CONSTRAINT

In addition to the channel conditions, in this work we consider a QoS-aware numerology and resource allocation by imposing LL constraints on LLUs. A LL transmission is characterized by a LL parameter τ that does not explicitly show up in the optimization problem defined in (8), but is is a part of the optimization problem through a parameter C_{LL}. As already discussed and shown in Fig. 4, C_{LL} predefines the total number of resources that can be assigned to LLUs.

Moreover, our optimization specifies a lower bound on the achievable rate of LLUs, which is related to transmission of the packet size q:

$$\sum_{\eta\in\mathcal{W}} \mathrm{R}_{\Delta f}^{u_{\mathrm{LL}}} N_{\Delta f}^{u_{\mathrm{LL}},\eta}\eta \geq q, \quad \forall u_{\mathrm{LL}} \in \mathcal{U}_{\mathrm{LL}}. \tag{9}$$

Therefore, the base station allocates as many resources to LLUs as they need in order to achieve q within the latency constraint τ. By imposing (9), we impact the total number of resources that can be assigned to NLLUs. It means that a smaller packet size that has to be transmitted by LLUs leaves more available resources for NLLUs, thereby affecting their achievable rate, and vice versa.

3) ACTIVITY CONSTRAINT

Only blocks with one unique mini slot length can be assigned to each LLU. Therefore, multiplexing of different mini-slots for a single user is not allowed. This constraint facilitates scheduling and simplify the optimization problem by controlling the positions of the resources of LLUs/NLLUs. It is given by means of the binary optimization variable $a_{\Delta f}^{u_{\mathrm{LL}},\eta}$:

$$\sum_{\eta\in\mathcal{W}} a_{\Delta f}^{u_{\mathrm{LL}},\eta} = 1, \quad a_{\Delta f}^{u_{\mathrm{LL}},\eta} \in \{0,1\}. \tag{10}$$

NLLUs are not restricted to the blocks with specific mini-slot durations. Since we want to maximize the achievable rate of NLLUs, they can select multiple mini-slot lengths. The binary variable of a NLLU is denoted by $a_{\Delta f}^{u_{\mathrm{NLL}},\eta}$.

Activity variables $a_{\Delta f}^{u_{\mathrm{LL}},\eta}$ and $a_{\Delta f}^{u_{\mathrm{NLL}},\eta}$ are functions of the variables $N_{\Delta f}^{u_{\mathrm{LL}},\eta}$ and $N_{\Delta f}^{u_{\mathrm{NLL}},\eta}$, respectively:

$$a_{\Delta f}^{u_{\mathrm{LL}},\eta} \leq N_{\Delta f}^{u_{\mathrm{LL}},\eta} \leq a_{\Delta f}^{u_{\mathrm{LL}},\eta} N_{\max}^{\eta}, \quad \eta \in \mathcal{W}, \tag{11a}$$

$$a_{\Delta f}^{u_{\mathrm{NLL}},\eta} \leq N_{\Delta f}^{u_{\mathrm{NLL}},\eta} \leq a_{\Delta f}^{u_{\mathrm{NLL}},\eta} N_{\max}^{\eta}, \quad \eta \in \mathcal{W}. \tag{11b}$$

Hence, by deciding on the number of assigned blocks, the optimizer decides on the activity binary variables, i.e., if there are resources assigned to a specific user, the activity variable takes the value 1; otherwise, the activity variable takes the value 0. The upper bounds of the total number of mini-slots are determined by the maximum number of blocks of weight η that can be fit into the entire resource grid of a subframe, $N_{\max}^{\eta}$.

4) KNAPSACK CAPACITY CONSTRAINTS

Our optimization problem (6a) can be identified with a multiscenario max-min *knapsack* problem as a generalization of the standard *knapsack* problem [43].

C_{total} is an upper bound on the total number of resources occupied by both, LLUs and NLLUs:

$$\sum_{u_{\mathrm{LL}}\in\mathcal{U}_{\mathrm{LL}}} \sum_{\eta\in\mathcal{W}} N_{\Delta f}^{u_{\mathrm{LL}},\eta}\eta + \sum_{u_{\mathrm{NLL}}\in\mathcal{U}_{\mathrm{NLL}}} \sum_{\eta\in\mathcal{W}} N_{\Delta f}^{u_{\mathrm{NLL}},\eta}\eta \leq C_{\mathrm{total}}. \tag{12}$$

5) UNIQUELY DEFINED RESOURCE CONSTRAINT

In cases when C_{LL} is not fully utilized by LLUs, we can assign those remaining resources to NLLUs. We distinguish two groups of NLLUs - the first group that shares the resources with LLUs and the second group which occupies the remaining resources:

$$\sum_{u_{\mathrm{LL}}\in\mathcal{U}_{\mathrm{LL}}} \sum_{\eta\in\mathcal{W}} N_{\Delta f}^{u_{\mathrm{LL}},\eta}\eta + \sum_{u_{\mathrm{NLL}}^{I}\in\mathcal{U}_{\mathrm{NLL}}^{I}} \sum_{\eta\in\mathcal{W}} N_{\Delta f}^{u_{\mathrm{NLL}}^{I},\eta}\eta = C_{\mathrm{LL}}, \tag{13a}$$

$$\sum_{u_{\mathrm{NLL}}^{II}\in\mathcal{U}_{\mathrm{NLL}}^{II}} N_{\Delta f}^{u_{\mathrm{NLL}}^{II},\eta}\eta = C_{\mathrm{total}} - C_{\mathrm{LL}}. \tag{13b}$$

The total number of resources assigned to NLLUs has to correspond to the sum of the resources occupied by the first group of NLLUs and resources occupied by the second group of NLLUs:

$$\sum_{u_{\mathrm{NLL}}^{I}\in\mathcal{U}_{\mathrm{NLL}}^{I}} N_{\Delta f}^{u_{\mathrm{NLL}}^{I},\eta}\eta + \sum_{u_{\mathrm{NLL}}^{II}\in\mathcal{U}_{\mathrm{NLL}}^{II}} N_{\Delta f}^{u_{\mathrm{NLL}}^{II},\eta}\eta = \sum_{u_{\mathrm{NLL}}\in\mathcal{U}_{\mathrm{NLL}}} N_{\Delta f}^{u_{\mathrm{NLL}},\eta}\eta. \tag{14}$$

A single NLLU can be simultaneously scheduled on resources assigned prior and after the latency constraint, so that one part of user resources belongs to the first group and the remaining resources belong to the second group of NLLUs.

B. MIXED NUMEROLOGY METHOD

1) PROBLEM FORMULATION

Similarly to the mini-slot approach, the optimization problem for the mixed numerology is composed of three parts: the objective function defined in (15a), optimization variables defined in (15b) and optimization constraints defined in (15c). It is given by:

$$\mathbf{P3}: \max \min_{u_{\mathrm{NLL}}\in\mathcal{U}_{\mathrm{NLL}}} \mathrm{R}_{\Delta f_s}^{u_{\mathrm{NLL}}} N_{\Delta f_s}^{u_{\mathrm{NLL}}}\eta \tag{15a}$$

$$\begin{aligned}
&\text{w.r.t.} \Delta f_s \in \mathcal{S}, N_{\Delta f_s}^{u_{\mathrm{NLL}}} \in \mathbb{N}_0, N_{\Delta f_s}^{u_{\mathrm{LL}}} \in \mathbb{N}_0,\\
&\quad N_{\Delta f_s}^{u_{\mathrm{NLL}}^{I}} \in \mathbb{N}_0, N_{\Delta f_s}^{u_{\mathrm{NLL}}^{II}} \in \mathbb{N}_0, K_{\Delta f_s} \in \mathbb{N}_0\\
&\quad a_{\Delta f_s}^{u_{\mathrm{NLL}}} \in \{0,1\}, a_{\Delta f_s}^{u_{\mathrm{LL}}} \in \{0,1\}, a_{\Delta f_s}^{\mathrm{gl}} \in \{0,1\},
\end{aligned} \tag{15b}$$

$$\begin{aligned}
&\text{subject to } \mathrm{R}_{\Delta f_s}^{u_{\mathrm{LL}}} N_{\Delta f_s}^{u_{\mathrm{LL}}}\eta \geq q, \forall u_{\mathrm{LL}} \in \mathcal{U}_{\mathrm{LL}},\\
&\quad \sum_{\Delta f_s\in\mathcal{S}} a_{\Delta f_s}^{u_{\mathrm{LL}}} = 1,\\
&\quad \sum_{\Delta f_s\in\mathcal{S}} a_{\Delta f_s}^{u_{\mathrm{NLL}}} = 1,\\
&\quad a_{\Delta f_s}^{u_{\mathrm{LL}}} \leq N_{\Delta f_s}^{u_{\mathrm{LL}}} \leq a_{\Delta f_s}^{u_{\mathrm{LL}}} N_{\max}^{\eta=14}, \forall \Delta f_s \in \mathcal{S},\\
&\quad a_{\Delta f_s}^{u_{\mathrm{NLL}}} \leq N_{\Delta f_s}^{u_{\mathrm{NLL}}} \leq a_{\Delta f_s}^{u_{\mathrm{NLL}}} N_{\max}^{\eta=14}, \forall \Delta f_s \in \mathcal{S},\\
&\quad a_{\Delta f_s}^{\mathrm{gl}} \leq \sum_{u_{\mathrm{NLL}}\in\mathcal{U}_{\mathrm{NLL}}} a_{\Delta f_s}^{u_{\mathrm{NLL}}}\\
&\quad + \sum_{u_{\mathrm{LL}}\in\mathcal{U}_{\mathrm{LL}}} a_{\Delta f_s}^{u_{\mathrm{LL}}} \leq a_{\Delta f_s}^{\mathrm{gl}} N_{\mathrm{users}},
\end{aligned}$$

$$\sum_{u_{\mathrm{LL}}\in\mathcal{U}_{\mathrm{LL}}}\sum_{\Delta f_s\in\mathcal{S}} N^{u_{\mathrm{LL}}}_{\Delta f_s} + \sum_{u_{\mathrm{NLL}}\in\mathcal{U}_{\mathrm{NLL}}}\sum_{\Delta f_s\in\mathcal{S}} N^{u_{\mathrm{NLL}}}_{\Delta f_s} + \eta G\Big(\sum_{\Delta f_s} a^{\mathrm{gl}}_{\Delta f_s} - 1\Big) \le C_{\mathrm{total}},$$

$$\sum_{\Delta f_s} K_{\Delta f_s} 2^{s-1} \le K_{15\,\mathrm{kHz}} - \sum_{\Delta f_s} a^{\mathrm{gl}}_{\Delta f_s} G + G,$$

$$\sum_{u_{\mathrm{LL}}\in\mathcal{U}^{s}_{\mathrm{LL}}} N^{u_{\mathrm{LL}}}_{\Delta f_s} + \sum_{u^{I}_{\mathrm{NLL}}\in\mathcal{U}^{I,s}_{\mathrm{NLL}}} N^{u^{I}_{\mathrm{NLL}}}_{\Delta f_s} = M^{\tau}_{\Delta f_s} K_{\Delta f_s},$$

$$\sum_{u^{II}_{\mathrm{NLL}}\in\mathcal{U}^{II,s}_{\mathrm{NLL}}} N^{u^{II}_{\mathrm{NLL}}}_{\Delta f_s} = \Big(M^{1ms}_{\Delta f_s} - M^{\tau}_{\Delta f_s}\Big) K_{\Delta f_s},$$

$$\sum_{u^{I}_{\mathrm{NLL}}\in\mathcal{U}^{I,s}_{\mathrm{NLL}}} N^{u^{I}_{\mathrm{NLL}}}_{\Delta f_s\, u^{I}_{\mathrm{NLL}}} + \sum_{u^{II}_{\mathrm{NLL}}\in\mathcal{U}^{II,s}_{\mathrm{NLL}}} N^{u^{II}_{\mathrm{NLL}}}_{\Delta f_s\, u^{II}_{\mathrm{NLL}}} = \sum_{u_{\mathrm{NLL}}\in\mathcal{U}^{s}_{\mathrm{NLL}}} N^{u_{\mathrm{NLL}}}_{\Delta f_s}, \tag{15c}$$

with $\eta = 14$ being the unique weight for all subcarrier spacings and for all users. Unlike the mini-slot approach, which employs a single subcarrier spacing, the resource allocation within the mixed numerology approach is based on the choice of a suitable subcarrier spacing for each user. Hence, in addition to the optimal number of blocks occupied by LLUs $N^{u_{\mathrm{LL}}}_{\Delta f_s}$ and optimal number of blocks occupied by NLLUs $N^{u_{\mathrm{NLL}}}_{\Delta f_s}$, we investigate the optimal numerology for each LLU and NLLU, i.e. subcarrier spacings Δf_s, where $s = \{1, 2, 3, 4\}$ corresponds to $S = \{15\,\mathrm{kHz}, 30\,\mathrm{kHz}, 60\,\mathrm{kHz}\ 120\,\mathrm{kHz}\}$, respectively. In order to obtain the total number of REs, these blocks have to be multiplied with weight $\eta = 14$, regardless of subcarrier spacing, as emphasized in Fig. 3b. The maximum number of blocks with weight $\eta = 14$ is denoted by $N^{\eta=14}_{\max}$. The activity of corresponding numerology is denoted by the binary optimization variable $a^{\mathrm{gl}}_{\Delta f_s}$. The binary optimization variables $a^{u_{\mathrm{NLL}}}_{\Delta f_s}$ and $a^{u_{\mathrm{LL}}}_{\Delta f_s}$ denote the activity of numerology occupied by a NLLU and a LLU, respectively. The total number of resources is divided not only among the total number of blocks occupied by NLLUs/LLUs, but also among the guard band G between numerologies. The total number of subcarriers within a subband bandwidth is denoted by $K_{\Delta f_s}$ and the total number of subcarriers of the 15 kHz subcarrier spacing within the available bandwidth is denoted by $K_{15\,\mathrm{kHz}}$. Similarly to the mini-slot approach, we distinguish two groups of users: the first group $\mathcal{U}^{I,s}_{\mathrm{NLL}}$ which shares the resources with LLUs within the total capacity of LLUs C_{LL}, and the second group $\mathcal{U}^{II,s}_{\mathrm{NLL}}$ which occupies the remaining resources. The total number of blocks per user from the first group of NLLUs is denoted by $N^{u^{I}_{\mathrm{NLL}}}_{\Delta f_s}$ and the total number of blocks per user from the second group of NLLUs is denoted by $N^{u^{II}_{\mathrm{NLL}}}_{\Delta f_s}$.

Following the idea of (8), the problem given in (15) can be reformulated as:

$$\mathbf{P4}: \quad \max\ z \quad \text{w.r.t. (15b)} \quad \text{subject to} \sum_{\Delta f_s\in\mathcal{S}} \mathrm{R}^{u_{\mathrm{NLL}}}_{\Delta f_s} N^{u_{\mathrm{NLL}}}_{\Delta f_s}\eta \ge z, \quad \forall u_{\mathrm{NLL}} \in \mathcal{U}_{\mathrm{NLL}} \quad \text{(15c)}. \tag{16}$$

2) LATENCY CONSTRAINT

Similarly to III-A2, we impose the latency constraint within the mixed numerology approach:

$$\mathrm{R}^{u_{\mathrm{LL}}}_{\Delta f_s} N^{u_{\mathrm{LL}}}_{\Delta f_s}\eta \ge q, \quad \forall u_{\mathrm{LL}} \in \mathcal{U}_{\mathrm{LL}}. \tag{17}$$

It is important to mention that in this case a LL parameter τ impacts the number of numerologies that the base station can assign to users. The smaller this parameter, the narrower the set of possible numerologies.

3) ACTIVITY CONSTRAINTS

Within the mixed numerology approach, users can choose between all possible subcarrier spacings from set $\mathcal{S}$, but each LLU/NLLU can select only one optimal subcarrier spacing. In such a way, we consolidate all resources of a specific user within a single subcarrier spacing. The corresponding constraints are given by means of the binary optimization variables $a^{u_{\mathrm{LL}}}_{\Delta f_s}$ for LLUs and $a^{u_{\mathrm{NLL}}}_{\Delta f_s}$ for NLLUs:

$$\sum_{\Delta f_s\in\mathcal{S}} a^{u_{\mathrm{LL}}}_{\Delta f_s} = 1, \quad a^{u_{\mathrm{LL}}}_{\Delta f_s} \in \{0, 1\} \tag{18a}$$

$$\sum_{\Delta f_s\in\mathcal{S}} a^{u_{\mathrm{NLL}}}_{\Delta f_s} = 1, \quad a^{u_{\mathrm{NLL}}}_{\Delta f_s} \in \{0, 1\}. \tag{18b}$$

These binary variables show the activity status of corresponding LLU/NLLU within subcarrier spacing Δf_s. If the user is active, this variable takes on the value one, otherwise it is zero.

Similarly to the mini-slot approach, these activity variables are determined by the number of assigned blocks to LLUs/NLLUs:

$$a^{u_{\mathrm{LL}}}_{\Delta f_s} \le N^{u_{\mathrm{LL}}}_{\Delta f_s} \le a^{u_{\mathrm{LL}}}_{\Delta f_s} N^{\eta=14}_{\max}, \quad \forall \Delta f_s \in \mathcal{S}, \tag{19a}$$

$$a^{u_{\mathrm{NLL}}}_{\Delta f_s} \le N^{u_{\mathrm{NLL}}}_{\Delta f_s} \le a^{u_{\mathrm{NLL}}}_{\Delta f_s} N^{\eta=14}_{\max}, \quad \forall \Delta f_s \in \mathcal{S}. \tag{19b}$$

In addition to the user activity variable, we define the binary variable $a^{\mathrm{gl}}_{\Delta f_s}$ that refers to the activity status of subcarrier spacing Δf_s. More precisely, this global binary variable indicates whether a specific subcarrier spacing is occupied by any of the users (LLUs and NLLUs) or not. Therefore, it is not a characteristic of a single LLU/NLLU, but rather a characteristic of a subband that might contain multiple LLUs and/or NLLUs. Thus, the global indicator variable $a^{\mathrm{gl}}_{\Delta f_s}$ of a given subcarrier spacing Δf_s has to be one if any user occupies this subcarrier spacing and zero otherwise:

$$a^{\mathrm{gl}}_{\Delta f_s} \le \sum_{u_{\mathrm{NLL}}\in\mathcal{U}_{\mathrm{NLL}}} a^{u_{\mathrm{NLL}}}_{\Delta f_s} + \sum_{u_{\mathrm{LL}}\in\mathcal{U}_{\mathrm{LL}}} a^{u_{\mathrm{LL}}}_{\Delta f_s} \le a^{\mathrm{gl}}_{\Delta f_s} N_{\mathrm{users}}. \tag{20}$$

4) KNAPSACK CAPACITY CONSTRAINTS

Below we define an upper bound on the available resources for both groups of users within a given bandwidth in the frequency domain and 1 ms of time span.

Having a constant weight η for different numerologies implies that the same number REs can be placed within a total capacity C_{total}. This allows to define C_{total} and C_{LL} in the same way as in (7). The total capacity does not only include the resources occupied by LLUs and the resources occupied by NLLUs, but also the guard band between different numerologies, G:

$$\sum_{u_{\text{LL}}\in\mathcal{U}_{\text{LL}}}\sum_{\Delta f_s\in\mathcal{S}} N_{\Delta f_s}^{u_{\text{LL}}} + \sum_{u_{\text{NLL}}\in\mathcal{U}_{\text{NLL}}}\sum_{\Delta f_s\in\mathcal{S}} N_{\Delta f_s}^{u_{\text{NLL}}} + \eta G\Big(\sum_{\Delta f_s} a_{\Delta f_s}^{\text{gl}} - 1\Big) \leq C_{\text{total}}. \tag{21}$$

Based on the threshold that we set as an upper bound on allowed INI in Sec. II-A and Sec. II-B, we determine the size of the guard band with respect to the 15 kHz subcarrier spacing in the frequency domain.

5) UNIQUELY DEFINED RESOURCE CONSTRAINT

This constraint checks that the total assigned bandwidth does not exceed the total available bandwidth. To achieve this, we introduce the following constraint:

$$\sum_{\Delta f_s} K_{\Delta f_s} 2^{s-1} \leq K_{15\,\text{kHz}} - \sum_{\Delta f_s} a_{\Delta f_s}^{\text{gl}} G + G, \tag{22}$$

where $K_{\Delta f_s}$ is the total number of subcarriers within subband s, $K_{15\,\text{kHz}}$ is the total number of subcarriers of the 15 kHz subcarrier spacing within the available bandwidth and $a_{\Delta f_s}^{\text{gl}}$ is the binary variable that shows the activity of corresponding numerology Δf_s. In our case, $K_{\Delta f_s}$ is an optimization variable. Since the frequency domain allocation is performed with respect to 15 kHz, the number of subcarriers $K_{\Delta f_s}$ has to be scaled by a factor 2^{s-1}.

It can happen that not all available resources within C_{LL} are occupied by LLUs and therefore, we assign the remaining resources to NLLUs. Again, we distinguish two groups of NLLUs: the first group $\mathcal{U}_{\text{NLL}}^{I}$ which shares the resources with LLUs within C_{LL}, and the second group $\mathcal{U}_{\text{NLL}}^{II}$ which occupies the remaining resources. The following equations are given with respect to the subset of LLUs/NLLUs that are assigned to a specific subband s. We denote the set of all LLUs which correspond to the same subband s by $\mathcal{U}_{\text{LL}}^{s}$. Similarly, we denote sets of NLLUs from the first and second group and the same subband by $\mathcal{U}_{\text{NLL}}^{I,s}$ and $\mathcal{U}_{\text{NLL}}^{II,s}$, respectively. For the resources that lie below τ, we define the following constraint:

$$\sum_{u_{\text{LL}}\in\mathcal{U}_{\text{LL}}^{s}} N_{\Delta f_s}^{u_{\text{LL}}} + \sum_{u_{\text{NLL}}^{I}\in\mathcal{U}_{\text{NLL}}^{I,s}} N_{\Delta f_s}^{u_{\text{NLL}}^{I}} = M_{\Delta f_s}^{\tau} K_{\Delta f_s},\ \forall \Delta f_s \in \mathcal{S}. \tag{23}$$

For the remaining resources we define a similar constraint:

$$\sum_{u_{\text{NLL}}^{II}\in\mathcal{U}_{\text{NLL}}^{II,s}} N_{\Delta f_s}^{u_{\text{NLL}}^{II}} = \Big(M_{\Delta f_s}^{1ms} - M_{\Delta f_s}^{\tau}\Big) K_{\Delta f_s}. \tag{24}$$

Additionally, the sum of resources from the first group of NLLUs and resources from the second group of NLLUs at appropriate subcarrier spacing has to correspond to the total number of resources of NLLUs:

$$\sum_{u_{\text{NLL}}^{I}\in\mathcal{U}_{\text{NLL}}^{I,s}} N_{\Delta f_s^{I}}^{u_{\text{NLL}}^{I}} + \sum_{u_{\text{NLL}}^{II}\in\mathcal{U}_{\text{NLL}}^{II,s}} N_{\Delta f_s^{II}}^{u_{\text{NLL}}^{II}} = \sum_{u_{\text{NLL}}\in\mathcal{U}_{\text{NLL}}^{s}} N_{\Delta f_s}^{u_{\text{NLL}}}. \tag{25}$$

IV. COMPUTATIONAL COMPLEXITY REDUCTION

So far, we have shown how we define our optimization problems. We can solve these optimization problems using any ILP solver, such as *intlinprog* from the Matlab optimization toolbox. However, ILP is NP-hard. Additionally to this optimizer, in this work we apply the *Dantzig-Wolfe* decomposition method followed by a *column generation* algorithm. By enabling a parallel execution of the set of mutually independent linear programming problems, the Dantzig-Wolfe decomposition provides the optimal solution with reduced computational complexity of the algorithm for large-scale scenarios compared to the two previously explained approaches [44].

A. OVERVIEW OF THE DANTZIG-WOLFE DECOMPOSITION

The *Dantzig-Wolfe* algorithm decomposes the original problem into a number of pricing or subproblems and a master problem that couples and coordinates the subproblems [45]. The idea of decomposition into subproblems is to obtain a solver-friendly form. This is achieved by carrying out the optimization on a smaller set of variables compared to the ILP techniques, where the size of the original problem is larger, implying a more computationally expensive execution. In order to employ the *Dantzig-Wolfe* algorithm, the constraints of the optimization problem have to be decomposable in a primal block angular structure. In our specific case, this can be achieved by partitioning the constraints into those relating to NLLUs and those relating to LLUs, as shown in Fig. 5. These subblocks are formed from the constraints that depend only on the optimization variables either connected to NLLUs or to LLUs. On top of these blocks, we accommodate linking (master) constraints that couple the optimization variables of both NLLUs and LLUs. To clarify this, we consider an example of the optimization method for the mini-slot approach defined in Sec. III-A. A similar approach can be applied in the mixed numerology case. We split our constraints as follows:

1) The group of NLL constraints involves the first constraint of (8) that is relevant for NLLUs, as well as, (11b), (13b) and (14).

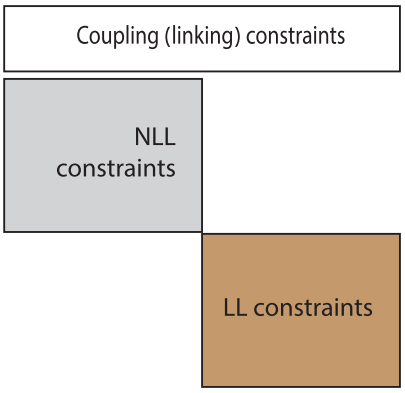


FIGURE 5. Primal block angular structure of the constraint matrix.

2) The group of LL constraints involves (9), (10) and (11a).
3) The linking constraints are defined in (12) and (13a).

V. SIMULATION RESULTS

In this section, we exhibit the performance of our optimization methods by numerical simulations. We assume a multi-user scenario of $N_{\text{users}} = 20$, $N_{\text{LLU}} = 10$ and $N_{\text{NLLU}} = 10$. Each user experiences different channel conditions characterized by the Doppler and delay spread of the channel.

The simulation parameters are summarized in Table 2.

TABLE 2. Simulation parameters.

s	$s=1$	$s=2$	$s=3$	$s=4$
Subcarrier spacing [kHz]	15	30	60	120
Number of subcarriers	576	288	144	72
Number of resource elements	8064			
Bandwidth [MHz]	8.64			
RMS delay spread [ns]	TDL-A $\{80, 180, 280, 380, 480\}$			
Velocity [km/h]	$\{5, 50, 100, 200, 300\}$			
Doppler spread [Hz]	$\{27.3, 273.3, 546.7, 1093.3, 1640\}$			
Number of users	10 LLUs + 10 NLLUs			
Number of realization	20			
τ [ms]	$1, 0.25$			
Waveform	OFDM, UFMC			
Carrier frequency [GHz]	5.9			

We show the results for three different scenarios with respect to the channel conditions employing either OFDM or UFMC. Therefore, the waveform employed by a base station is unique for all numerologies and it does not vary between them. The first scenario (Scenario I) assumes a fixed RMS delay spread of 200 ns for each user and various Doppler spreads characterized by the velocities from set $\{5, 50, 100, 200, 300\}$ km/h. We randomly assign a velocity from this set to each user and the optimization results are averaged over 20 random realizations. The second scenario (Scenario II) assumes a fixed Doppler spread of 1.093 kHz, (200 km/h velocity) and Jakes' Doppler spectrum, and various RMS delay spreads from set $\{80, 180, 280, 380, 480\}$ ns. The third scenario (Scenario III) assumes opposing channel conditions: one group of users experiences large RMS delay spreads from set $\{380, 480\}$ ns and small Doppler spread of 273.337 Hz (50 km/h velocity), whereas the other group of users experiences large Doppler spread of 1.093 kHz (200 km/h velocity) and RMS delay spread from set $\{80, 180\}$ ns. We use a TDL-A channel model for all scenarios and employ the Vienna 5G link level simulator to generate the channel traces, [46], [47]. The number of subcarriers, and therefore the bandwidth, was chosen such that it provides meaningful results in terms of the achievable rate on the one hand and that the solution is obtained with reasonable amount of computing effort on the other hand [48]. The total number of resource elements is calculated as a multiplication of the total number of available subcarriers of corresponding subcarrier spacing and the total number of symbols of corresponding duration within 1 ms, according to (7) and Table 2. The number of users was selected according to the bandwidth size and packet size q, so that all users can be supported within the available bandwidth and the maximum considered packet size.

In the following, we compare the results imposing different latency constraints τ. Specifically, we select latency values τ that are equal to the slot durations of 15 kHz and 60 kHz, i.e., 1 ms and 0.25 ms, respectively. The results are provided with respect to two metrics: percentage of served LLUs (α) and sum achievable rate of NLLUs. The LLUs occupy as many resources as they need to convey their packet size q within the latency constraint τ. The remaining resources are assigned to NLLUs. Correspondingly, the value of the packet size q has an impact on the sum achievable rate of NLLUs. Larger q implies more resources for LLUs and hence less resources for NLLUs. However, if q is too large such that we cannot satisfy all LLUs within capacity C_{LL}, then some of the LLUs cannot be served within their latency constraint τ and we assume that the scheduler then discards them; these users are therefore in outage. In order to calculate the percentage of served LLUs, we define a metric α as:

$$\alpha[\%] = \frac{\hat{N}_{\text{LLU}}}{N_{\text{LLU}}} 100, \tag{26}$$

with $\hat{N}_{\text{LLU}}$ being the number of satisfied LLUs. In case none of the LLUs can be satisfied, the optimization is infeasible. For the mini-slot approach we firstly discard the users with the lowest SINR, since they require most of the resources. We repeatedly discard LLUs, until we obtain a feasible solution of (8). For the mixed numerology approach, we discard users randomly, since different subcarrier spacings can be assigned to different users and it is therefore not possible to uniquely identify users with lowest SINRs (the SINR depends on the subcarrier spacing).

In Fig. 6 we show the sum achievable rate of NLLUs as a function of the packet size q of LLUs for $\tau = 1$ ms. In this case, due to the relaxed latency constraint, all LLUs are satisfied for all q and $\alpha = 100\,\%$. We observe that the different subcarrier spacings achieve very different

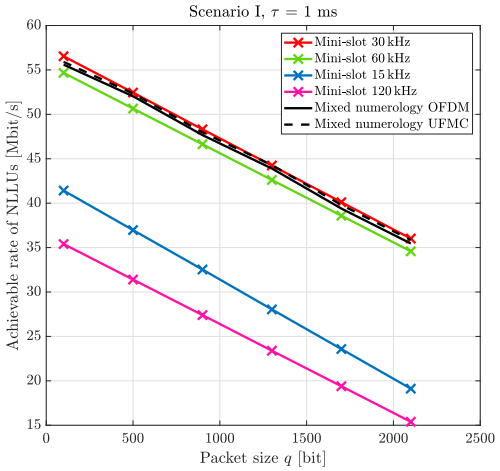


FIGURE 6. Sum achievable rate of NLLUs for Scenario I and $\tau = 1$ ms.

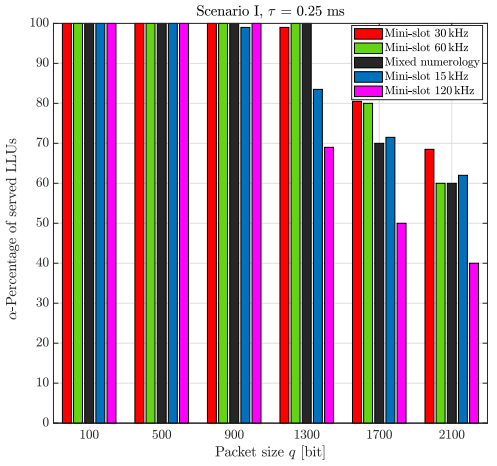


FIGURE 7. Percentage of served LLUs for Scenario I and $\tau = 0.25$ ms.

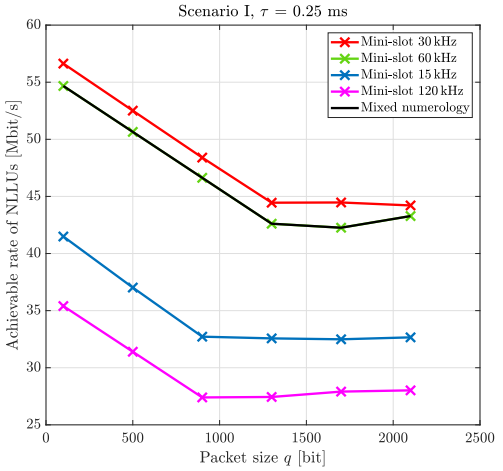


FIGURE 8. Sum achievable rate of NLLUs for Scenario I and $\tau = 0.25$ ms.

performance. The best performance in our Scenario I is achieved by 30 kHz, which provides on average the highest SINR for the employed set of channel characteristics. Subcarrier spacings of 15 kHz and 120 kHz perform worst in our Scenario I: 15 kHz exhibits strong ICI at larger velocities, whereas 120 kHz exhibits ISI for 200 ns of RMS delay spread. For the mini-slot approach, we have only simulated CP-OFDM, since with mini-slots there is no INI and therefore CP-OFDM and UFMC achieve the same performance. However, in the case of mixed numerology, two different subcarrier spacings that achieve the best SINR values (30 kHz and 60 kHz) are assigned to users. In this case, INI exists and some resources are assigned to the guard band. From Fig. 6 it can be observed that the mixed numerology method achieves similar performance as the mini-slot approach for 30 kHz, irrespective of the overhead caused by the guard bands. Employing UFMC in the mixed numerology case provides only a minor gain. This gain is due to smaller guard bands, which are reduced from 12 subcarriers for OFDM to 8 subcarriers for UFMC. The reason for very similar guard overhead is that the interference of both systems close to the border between different numerologies is very similar as shown in Fig. 2.

In Fig. 7, we show the percentage of served LLUs for $\tau = 0.25$ ms under Scenario I. We observe that in this case not all LLUs can be served within the given LL capacity C_{LL} with growing packet size q; therefore, the scheduler has to discard some of these users, decreasing α. In addition, α is also impacted by the SINRs of the users; lower SINR implies smaller α and higher outage probability. Interestingly, the best performance for LLUs is not achieved by 120 kHz subcarrier spacing, as one could expect, simply because the SINR values achieved by employing this subcarrier spacing are lower compared to 60 kHz subcarrier spacing due to the imposed channel conditions. Hence, we notice that this subcarrier spacing experiences the lowest level of served LLUs.

In Fig. 8, we show the sum achievable rate for $\tau = 0.25$ ms under Scenario I. The order of the curves matches the one in Fig. 6. Unlike in Fig. 6 where curves have a linear behaviour, here we notice that after a certain packet size, the sum achievable rate saturates. The reason for this is that from a certain q on, all resources that are predefined for LLUs, C_{LL}, are fully occupied and therefore the remaining resources assigned to NLLUs stay constant with small variations. Additionally, in Fig. 8 we show the results obtained applying mixed numerology, employing both OFDM and UFMC. The main difference between the mini-slot approach and the mixed numerology approach is that former employs one numerology with different slot durations (mini-slots) and latter employs multiple numerologies with their regular slot durations according to Table 1. Therefore, by considering $\tau = 0.25$ ms, it is not possible to employ 15 kHz and 30 kHz numerologies whose slot durations are 1 ms and 0.5 ms, respectively. Due to the narrow choice of subcarrier spacings (60 kHz and 120 kHz) that achieve $\tau = 0.25$ ms and poor performance of 120 kHz due to the considered RMS delay spread, the optimization provides the resource allocation based only on 60 kHz subcarrier spacing. In such a case, when only a single subcarrier spacing is employed, UFMC

provides the same result as OFDM. The reason for that is a lack of the guard band. Therefore, the mixed numerology case of both OFDM and UFMC and the mini-slot case with 60 kHz subcarrier spacing overlap.

The sum achievable rate of NLLUs under Scenario II with $\tau = 1$ ms, as well as, the percentage of served LLUs and the sum achievable rate of NLLUs under Scenario II with $\tau = 0.25$ ms are shown in Fig. 9, 10 and Fig. 11, respectively. In addition to the lower sum achievable rate of NLLUs compared to the same τ under Scenario I, we can also notice the saturation already with small packet sizes. This indicates that Scenario II is less suitable than Scenario I for the employed numerology. However, the order of the curves remains unchanged with an optimal $\Delta f = 30$ kHz in both cases.

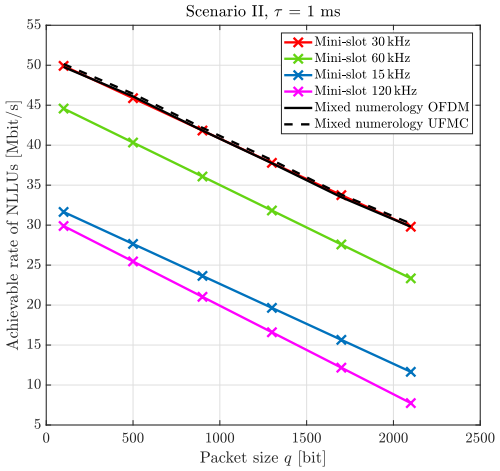


FIGURE 9. Sum achievable rate of NLLUs for Scenario II and $\tau = 1$ ms.

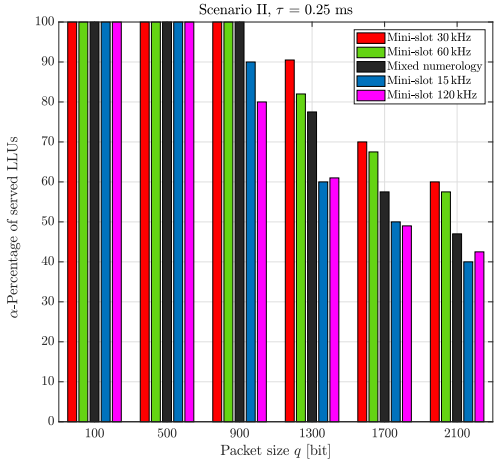


FIGURE 10. Percentage of served LLUs for Scenario II and $\tau = 0.25$ ms.

In Fig. 12, we show the sum achievable rate of NLLUs under Scenario III with the latency constraint $\tau = 1$ ms. In this case, mixed numerology shows better adjustment to different channel conditions and hence achieves better performance than a single numerology. The group of users that experience small Doppler spread and large RMS delay spread suffer from

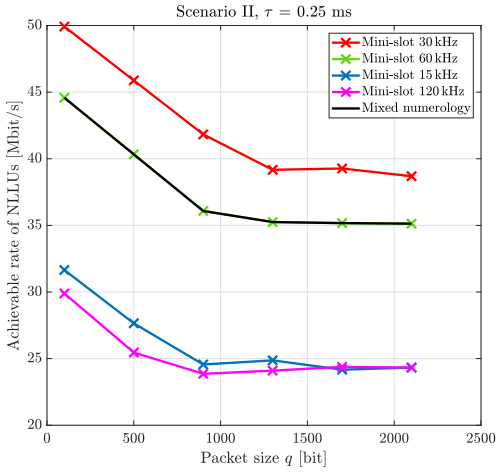


FIGURE 11. Sum achievable rate of NLLUs for Scenario II and $\tau = 0.25$ ms.

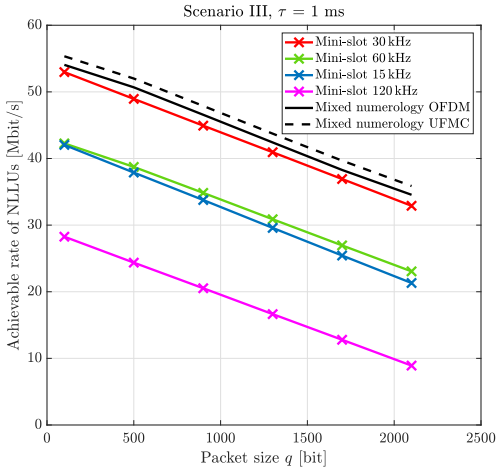


FIGURE 12. Sum achievable rate of NLLUs for Scenario III and $\tau = 1$ ms.

large ISI. Therefore, large subcarrier spacings are not suitable for this group of users. On the other hand, small subcarrier spacings are not convenient for large ICI caused by large Doppler spread for the other group of users. Additionally, we observe that 15 kHz subcarrier spacing exhibits larger achievable rate compared to Scenario I and Scenario II due to the large RMS delay spreads. Similarly to these scenarios, 30 kHz subcarrier spacing manifests the best characteristics among numerologies.

As we already discussed, the reliability aspect of uRLLC would require a sufficiently large fading margin to the SINR/achievable rate. In cases of a flexible latency constraint, this margin would not impact the optimality of numerologies. However, in cases with a stringent latency constraint, it is more difficult to guarantee reliability and this margin could affect performance.

In Fig. 13 we compare our results with the state of the art related work presented in [23] considering $\tau = 1$ ms. As proposed in [23], we consider two groups of users: the first group experiences ITU Vehicular-A channel parameters with

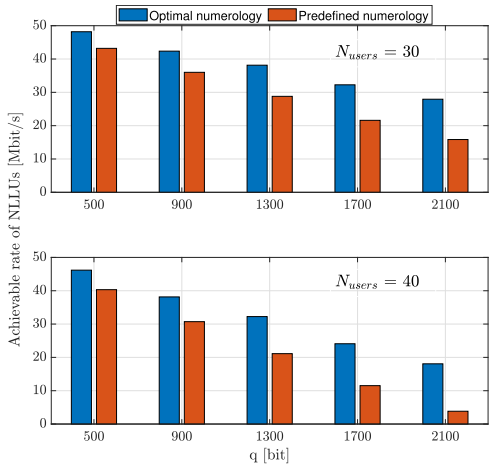


FIGURE 13. Comparison of our approach with the state of the art work with $\tau = 1$ ms.

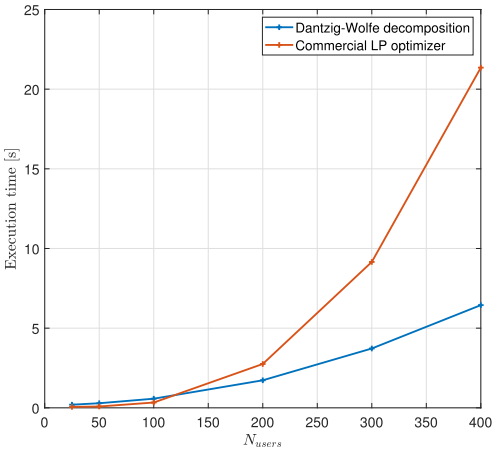


FIGURE 14. Comparison of the commercial LP optimizer vs. DW decomposition.

Doppler shift of 500 Hz, whereas the other group experiences ITU Pedestrian-B channel parameters with Doppler shift of 50 Hz. The authors predefine a specific numerology to each of the users based on their channel conditions and service demands. However, their approach is quite heuristic: for users that experience large Doppler spreads and require large data demands they adopt a numerology with the smallest subcarrier spacing, which corresponds to 15 kHz in our case. The other group comprises users with large RMS delay spreads and small velocities and they have average requirements on the transmission data. For this group of users the authors adopt a 30 kHz numerology. Our approach is based on selecting an optimal numerology based on the channel conditions and service demands. Given that 15 kHz numerology is not robust against ICI, i.e., at a high Doppler spread, the average SINR is very low and therefore it is not the best solution for high moving users. On the other hand, the 15 kHz numerology can be adopted to pedestrian users that experience a small Doppler and very high RMS delay spread, especially if their requirement is not a stringent latency constraint. Due to these reasons, we can conclude that the approach proposed in [23], is deficient compared to our approach for all packet sizes q.

In order to reduce computational complexity of the optimization for large-scale scenarios and obtain the optimal solution at the same time, in this work we investigate the importance of employing more sophisticated optimization methods such as the *Dantzing-Wolfe* decomposition. In Fig. 14 we compare the commercial LP[5] optimizer embedded in the Matlab optimization toolbox and the *Dantzing-Wolfe* decomposition followed by the *column generation* algorithm. The computational complexity of both solvers is the same and it is $\mathcal{O}(m^{2,38})$, where m is the input size of the optimization problem [49]. The gain visible in Fig. 14 arises due to the parallel execution of two independent sub-problems whose dimensions are smaller than m. The dimensions of the subproblems depend on the number of LLUs/NLLUs in our case. If we assume that a dimension m_1 of the first subproblem that is related to NLLUs and $m_1 < n$, then its complexity is given by $\mathcal{O}(m_1^{2,38})$ and similarly, for the second subproblem it is $\mathcal{O}(m_2^{2,38})$, where $m_2 < m$. It holds true that $\mathcal{O}(m_1^{2,38} + m_2^{2,38}) < \mathcal{O}((m_1 + m_2)^{2,38}) = \mathcal{O}(m^{2,38})$. Of course, the computational complexity within the Dantzig-Wolfe algorithm depends on the number of iterations that have to be executed in order to find an optimal solution. However, only one iteration is sufficient to achieve this in our case. The results are given with respect to the execution time of the optimization versus the number of users. We can observe that the *Dantzing-Wolfe* decomposition provides significantly better results for a large number of users due to the parallel execution on the set of optimization variables, rather than on the full-size problem. The commercial LP solvers are sophisticated enough for small number of users.

[5]Instead of an ILP solver that exhibits very high computational complexity, in this example we employ an LP solver that provides an approximate solution with smaller complexity.

VI. CONCLUSION

In order to support a number of different types of services and user requirements on the same carrier in 5G and beyond, two different methods have been proposed in the literature. In this work, we compare these two methods imposing a service demand in terms of latency to one priority group of users. Once we satisfy the priority users, our goal is to maximize the achievable rate for the remaining users. We formulate the constrained optimization problems for both methods. In addition to the latency constraint, users differ in SINRs, which, according to the results, has proven to be a dominant effect on performance for both methods in situations with strongly heterogeneous channel conditions in terms of delay and Doppler spread. In the case of more flexible latency constraints, yet not conflicting channel conditions both methods perform similarly. In cases when the channel conditions are quite opposite, for example in indoor scenarios on the one hand, and high-speed environment on the other hand,

the mixed numerology concept outperforms the mini-slot concept, since it pays off to employ more numerologies at the same time, even with a lack of resources assigned to the guard band. However, when the latency is stringent, the mixed numerology concept selects only the numerologies whose slot durations correspond to the latency constraint. Therefore, in such cases the mini-slot concept with the optimal subcarrier spacing outperforms the mixed numerology concept.

Employing a waveform such as UFMC, with better confined spectrum, slightly improves the performance due to a smaller amount of resources wasted for the guard band. In addition, we employ the *Dantzig-Wolfe* decomposition followed by a *column generation* algorithm for complexity reduction of the optimization execution in large-scale scenarios. Our comparison of execution time shows significant improvements over a standard general purpose solver.

In the future, it would be interesting to investigate a scenario that considers multiple numerologies in conjunction with mini-slots and to evaluate the optimal numerologies and resource allocations for different user and service requirements. Additionally, a study on efficient machine learning-based algorithms for supporting even greater flexibility in beyond 5G technologies should be conducted.

APPENDIX A
OVERVIEW OF PHY PARAMETERS

In the following we provide the closed-form expressions of ICI, ISI, INI and channel estimation error employing OFDM.

Assuming Jakes' Doppler spectrum, the ICI power can be calculated as [31]:

$$\sigma^{2,\mathrm{ofdm}}_{\mathrm{ICI}_{\Delta f}} = 1 - \int_{-1}^{1} (1-|x|) J_0(2\pi f_\mathrm{d} T^{\mathrm{ofdm}}_{\Delta f} x) dx, \quad (27)$$

with J_0 being the zeroth-order Bessel function of the first kind and $f_\mathrm{d} = \frac{v_{\max} f_\mathrm{c}}{c_0}$ being the maximum Doppler shift depending on the maximum user velocity $v_{\max}$, carrier frequency f_c and speed of light c_0. The OFDM symbol duration including CP is denoted by $T^{\mathrm{ofdm}}_{\Delta f} \approx 1/\Delta f$.

ISI depends on the power delay profile of the transmission channel $\rho(\xi)$, [32]:

$$\sigma^{2,\mathrm{ofdm}}_{\mathrm{ISI}_{\Delta f}} = \frac{1}{T^{\mathrm{ofdm}}_{\Delta f}} \int_0^{\xi_{\max} - T_{\mathrm{CP},\Delta f}} \int_{T_{\mathrm{CP},\Delta f}+t}^{\xi_{\max}} \rho(\xi) d\xi dt, \quad (28)$$

where $\xi_{\max}$ is the maximum delay.

The channel estimation error closed-form expression is provided in [50]:

$$\sigma^{2,\mathrm{ofdm}}_{\mathrm{e}_{\Delta f,k}} = c_{\mathrm{e},\Delta f} \frac{\sigma^{2,\mathrm{ofdm}}_{\mathrm{n}} + \sigma^{2,\mathrm{ofdm}}_{\mathrm{ICI}_{\Delta f}} + \sigma^{2,\mathrm{ofdm}}_{\mathrm{ISI}_{\Delta f}} + \sigma^{2,\mathrm{ofdm}}_{\mathrm{INI}_{\Delta f,k}}}{\sigma^2_{\mathrm{p}}} + d_{\mathrm{e},\Delta f}, \quad (29)$$

where $c_{\mathrm{e},\Delta f}$ is a constant that depends on the employed channel estimation method and $d_{\mathrm{e},\Delta f}$ is the interpolation error that additionally depends on the second order statistics of

TABLE 3. List of parameters used throughout the paper.

Parameter name	Notation
Numerology index	s
Set of all numerologies	$\mathcal{S}$
Received data symbol at frequency-time index k,n in OFDM	$Y^{\mathrm{ofdm}}_{k,n}$
Received data symbol at frequency-time index k,n in UFMC	$Y^{\mathrm{ufmc}}_{k,n}$
Transmitted data symbol at frequency-time index k,n in OFDM	$X^{\mathrm{ofdm}}_{k,n}$
Transmitted data symbol at frequency-time index k,n and subband s in UFMC	$X^{s}_{k,n}$
Frequency response of OFDM/UFMC	$H_{k,n}$
Interference and noise in OFDM	$W^{\mathrm{ofdm}}_{k,n}$
Interference and noise in UFMC	$W^{\mathrm{ufmc}}_{k,n}$
Frequency response of the time domain subband filter	$F^{s}_{k,n}$
Subcarrier spacing	Δf
Symbol duration	$T_{\Delta f}$
Cyclic prefix length	$T_{\mathrm{CP},\Delta f}$
Maximum Doppler shift	f_d
Maximum user velocity	$v_{\max}$
Power delay profile of the transmission channel	$\rho(\xi)$
Maximum delay	$\xi_{\max}$
Intercarrier interference power in OFDM	$\sigma^{2,\mathrm{ofdm}}_{\mathrm{ICI}_{\Delta f}}$
Intersymbol interference power in OFDM	$\sigma^{2,\mathrm{ofdm}}_{\mathrm{ISI}_{\Delta f}}$
Internumerology interference power in OFDM	$\sigma^{2,\mathrm{ofdm}}_{\mathrm{INI}_{\Delta f}}$
Channel estimation error in OFDM	$\sigma^{2,\mathrm{ofdm}}_{\mathrm{e}_{\Delta f}}$
Noise power in OFDM	$\sigma^{2,\mathrm{ofdm}}_{\mathrm{n}}$
Intercarrier interference power in UFMC	$\sigma^{2,\mathrm{ufmc}}_{\mathrm{ICI}_{\Delta f}}$
Intersymbol interference power in UFMC	$\sigma^{2,\mathrm{ufmc}}_{\mathrm{ISI}_{\Delta f}}$
Internumerology interference power in UFMC	$\sigma^{2,\mathrm{ufmc}}_{\mathrm{INI}_{\Delta f}}$
Channel estimation error in UFMC	$\sigma^{2,\mathrm{ufmc}}_{\mathrm{e}_{\Delta f}}$
Noise power in UFMC	$\sigma^{2,\mathrm{ofdm}}_{\mathrm{n}}$
Upper bound on allowed INI	$\sigma^{2}_{\mathrm{INI}^{\max}}$
Channel power	σ^2_H
Filter length	L
FFT length	N
Channel length	L_{CH}
User index	u
Low-latency user index	u_{LL}

TABLE 3. ***(Continued.)*** **List of parameters used throughout the paper.**

Parameter name	Notation
Non low-latency user index	u_{NLL}
Non low-latency user index from the first group of all non low-latency users	u^{I}_{NLL}
Non low-latency user index from the second group of all non low-latency users	u^{II}_{NLL}
Set of all users	$\mathcal{U}$
Set of all low-latency users	$\mathcal{U}_{\mathrm{LL}}$
Set of all non low-latency users	$\mathcal{U}_{\mathrm{NLL}}$
Set of all non low-latency users from the first group of NLLUs	$\mathcal{U}^{I}_{\mathrm{NLL}}$
Set of all non low-latency users from the second group of NLLUs	$\mathcal{U}^{II}_{\mathrm{NLL}}$
User achievable rate	$\mathrm{R}^{u}_{\Delta f}$
User SINR	$\bar{\gamma}^{u}_{\Delta f}$
Total number of low-latency users	N_{LL}
Total number of non low-latency users	N_{NLL}
Total number of users	N_{users}
Latency constraint	τ
Packet size	q
Mini-slot weight (number of symbols that compose a mini-slot)	η
Set of all mini-slots	$\mathcal{W}$
Total number of mini-slots with weight η occupied by a low-latency user	$N^{u_{\mathrm{LL}},\eta}_{\Delta f}$
Total number of mini-slots with weight η occupied by a non low-latency user	$N^{u_{\mathrm{NLL}},\eta}_{\Delta f}$
Activity variable of a low-latency user	$a^{u_{\mathrm{LLL}},\eta}_{\Delta f}$
Activity variable of a non low-latency user	$a^{u_{\mathrm{NLL}},\eta}_{\Delta f}$
Total number of available resources	C_{total}
Total number of available resources predefined for low-latency users	C_{LL}
Total number of subcarriers	$K_{\Delta f}$
Total number of symbols within a subframe duration of 1 ms	$M^{1ms}_{\Delta f}$
Total number of symbols within τ	$M^{\tau}_{\Delta f}$
Lower bound on the achievable rate of non low-latency users	z
Total number of mini-slots of a specific user u^{I}_{NLL}	$N^{u^{I}_{\mathrm{NLL}},\eta}_{\Delta f}$
Total number of mini-slots of a specific user u^{II}_{NLL}	$N^{u^{II}_{\mathrm{NLL}},\eta}_{\Delta f}$

TABLE 3. ***(Continued.)*** **List of parameters used throughout the paper.**

Parameter name	Notation
Total number of blocks occupied by a low-latency user within numerology s	$N^{u_{\mathrm{LL}}}_{\Delta f_s}$
Total number of blocks occupied by a non low-latency user within numerology s	$N^{u_{\mathrm{NLL}}}_{\Delta f_s}$
Activity variable of a low-latency user within numerology s	$a^{u_{\mathrm{LL}}}_{\Delta f_s}$
Activity variable of a non low-latency user within numerology s	$a^{u_{\mathrm{NLL}}}_{\Delta f_s}$
Activity variable of numerology s	$a^{\mathrm{gl}}_{\Delta f_s}$
Maximum number of blocks with weight $\eta = 14$	$N^{\eta=14}_{\mathrm{max}}$
Number of subcarriers within numerology Δf_s	$K_{\Delta f_s}$
Set of all non low-latency users from the first group of non low-latency users within numerology s	$\mathcal{U}^{I,s}_{\mathrm{NLL}}$
Set of all non low-latency users from the second group of non low-latency users within numerology s	$\mathcal{U}^{II,s}_{\mathrm{NLL}}$
Total number of blocks of a specific user u^{I}_{NLL} within numerology s	$N^{u^{I}_{\mathrm{NLL}}}_{\Delta f_s}$
Total number of blocks of a specific user u^{II}_{NLL} within numerology s	$N^{u^{II}_{\mathrm{NLL}}}_{\Delta f_s}$
Guard band	G
Total number of satisfied low-latency users	$\hat{N}_{\mathrm{LLU}}$
Percentage of satisfied LLUs	α
Input size of the optimization problem	m
Input size of the first subproblem related to non low-latency users	m_1
Input size of the second subproblem related to low-latency users	m_2
Interpolation error	$d_{\mathrm{e},\Delta f}$
Weighting factor that depends on the employed channel estimation	$c_{\mathrm{e},\Delta f}$
Filter transfer function	f_k
Pilot power	σ^2_{p}

the channel represented by the autocorrelation matrix [33]. We assume an equal power distribution on data and pilots σ^2_{p} and the diamond-shaped pilot pattern of 5G NR. By introducing an upper bound on INI in the next paragraphs, we simplify (29) such that we obtain a subcarrier independent, single-value measure representing an upper bound on the channel estimation error.

As already mentioned, we consider a *victim* user on the one hand and an *aggressor* user on the other hand. Hence, the user with a smaller subcarrier spacing Δf_s is considered to be the *victim* user and the user with a larger subcarrier spacing $\Delta f_{s' \in \{2,3,4\}}$ is considered to be the *aggressor* user. In [34] we derive the closed-form expression of INI caused by aggressor users $\Delta f_{s'}$ onto the victim user Δf_s:

$$\sigma^{2,\text{ofdm}}_{\text{INI}_{\Delta f_s,k}} = \sum_{k'=0}^{N_{\Delta f_{s'}}-1} \left|\text{sinc}\left((k'+1)\frac{\pi}{2^{s-1}} + k\pi\right)\right|^2, \quad (30)$$

with k and k' being the subcarrier indices of the neighboring subbands Δf_s and $\Delta f_{s'}$, respectively. The number of subcarriers within corresponding aggressor subbands is denoted by $N_{\Delta f_{s'}}$. Additionally, in [34] we derive a closed-form expression of INI power in case UFMC is employed, $\sigma^{2,\text{ufmc}}_{\text{INI}_{\Delta f_s,k}}$, again with respect to a victim user:

$$\sigma^{2,\text{ufmc}}_{\text{INI}_{\Delta f_s,k}} = \sum_{k'=0}^{N_{\Delta f_{s'}}-1} \left|\text{sinc}\left((k'+1)\frac{\pi}{2^{s-1}} + k\pi\right)\right|^2 |f(k)|^2, \quad (31)$$

where

$$f_k = \text{sinc}\left(\frac{k\pi L_{\Delta f_{s'}}}{2^{s-1} N_{\Delta f_{s'}}}\right). \quad (32)$$

As we can notice, the closed-form expression is composed of two parts - a pure OFDM part as in (30) and the filter transfer function $|f_k|^2$. The filter transfer function is a *sinc* function with an argument that depends on $L_{\Delta f_{s'}}$ that is the filter length and $N_{\Delta f_{s'}}$ that is the FFT length of the *aggressor* user.

APPENDIX B
LIST OF PARAMETERS USED THROUGHOUT THE PAPER

In the following, we list all parameters used in this work in Table 3.

REFERENCES

[1] A. Zaidi, F. Athley, J. Medbo, U. Gustavsson, G. Durisi, and X. Chen, *5G Physical Layer: Principles, Models and Technology Components*. New York, NY, USA: Academic, 2018.
[2] A. A. Zaidi, R. Baldemair, H. Tullberg, H. Bjorkegren, L. Sundstrom, J. Medbo, C. Kilinc, and I. Da Silva, "Waveform and numerology to support 5G services and requirements," *IEEE Commun. Mag.*, vol. 54, no. 11, pp. 90–98, Nov. 2016.
[3] *Study on and Requirements for Next Generation Access Technologies*, document TR 38.913, 3rd Generation Partnership Project, May 2017.
[4] *M Series, IMT Vision–Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond*, document M.2083-0, ITU Recommendation, 2015.
[5] H. Ji, S. Park, J. Yeo, Y. Kim, J. Lee, and B. Shim, "Ultra-reliable and low-latency communications in 5G downlink: Physical layer aspects," *IEEE Wireless Commun.*, vol. 25, no. 3, pp. 124–130, Jun. 2018.
[6] *5G; Service Requirements for Next Generation New Services and Markets*, document TS 22.261, 3rd Generation Partnership Project, 2018.
[7] G. Durisi, T. Koch, and P. Popovski, "Toward massive, ultrareliable, and low-latency wireless communication with short packets," *Proc. IEEE*, vol. 104, no. 9, pp. 1711–1726, Sep. 2016.
[8] S. Schwarz, T. Philosof, and M. Rupp, "Signal processing challenges in cellular-assisted vehicular communications: Efforts and developments within 3GPP LTE and beyond," *IEEE Signal Process. Mag.*, vol. 34, no. 2, pp. 47–59, Mar. 2017.
[9] S. Schwarz and M. Rupp, "Society in motion: Challenges for LTE and beyond mobile communications," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 76–83, May 2016.
[10] N. Patriciello, S. Lagen, L. Giupponi, and B. Bojovic, "5G new radio numerologies and their impact on the End-To-End latency," in *Proc. IEEE 23rd Int. Workshop Comput. Aided Model. Design Commun. Links Netw. (CAMAD)*, Sep. 2018, pp. 1–6.
[11] C. Campolo, A. Molinaro, F. Romeo, A. Bazzi, and A. O. Berthet, "5G NR V2X: On the impact of a flexible numerology on the autonomous sidelink mode," in *Proc. IEEE 2nd 5G World Forum (5GWF)*, Sep. 2019, pp. 102–107.
[12] L. Marijanovic, S. Schwarz, and M. Rupp, "Optimal numerology in OFDM systems based on imperfect channel knowledge," in *Proc. IEEE 87th Veh. Technol. Conf. (VTC Spring)*, Jun. 2018, pp. 1–5.
[13] *Technical Specification Group Services and System Aspects; Release 15 Description; Summary of Rel-15 Work Items*, document TR 21.915, 3rd Generation Partnership Project, 2019.
[14] *5G; Study on New Radio (NR) Access Techology*, document TR 38.912, 2018.
[15] B. Chang, L. Zhang, L. Li, G. Zhao, and Z. Chen, "Optimizing resource allocation in URLLC for real-time wireless control systems," *IEEE Trans. Veh. Technol.*, vol. 68, no. 9, pp. 8916–8927, Sep. 2019.
[16] J. Tang, A. Shojaeifard, D. K. C. So, K.-K. Wong, and N. Zhao, "Energy efficiency optimization for CoMP-SWIPT heterogeneous networks," *IEEE Trans. Commun.*, vol. 66, no. 12, pp. 6368–6383, Dec. 2018.
[17] A. Yazar and H. Arslan, "A flexibility metric and optimization methods for mixed numerologies in 5G and beyond," *IEEE Access*, vol. 6, pp. 3755–3764, 2018.
[18] L. Marijanovic, S. Schwarz, and M. Rupp, "Optimal resource allocation with flexible numerology," in *Proc. IEEE Int. Conf. Commun. Syst. (ICCS)*, Dec. 2018, pp. 136–141.
[19] 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)*, May 2019, pp. 1–6.
[20] 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.
[21] 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.
[22] T. Levanen, Z. Li, J. Talvitie, M. Renfors, and M. Valkama, "Filtered OFDM based URLLC in 5G new radio: Principles and performance," in *Proc. IEEE Wireless Commun. Netw. Conf. Workshop (WCNCW)*, Apr. 2019, pp. 1–7.
[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] W. Sui, X. Chen, S. Zhang, Z. Jiang, and S. Xu, "Energy-efficient resource allocation with flexible frame structure for heterogeneous services," in *Proc. Int. Conf. Internet Things (iThings) IEEE Green Comput. Commun. (GreenCom) IEEE Cyber, Phys. Social Comput. (CPSCom) IEEE Smart Data (SmartData)*, Jul. 2019, pp. 749–755.
[25] A. Gonzalez, S. Kuhlmorgen, A. Festag, and G. Fettweis, "Resource allocation for block-based multi-carrier systems considering QoS requirements," in *Proc. GLOBECOM IEEE Global Commun. Conf.*, Dec. 2017, pp. 1–7.
[26] J. Zhang, X. Xu, K. Zhang, B. Zhang, X. Tao, and P. Zhang, "Machine learning based flexible transmission time interval scheduling for eMBB and uRLLC coexistence scenario," *IEEE Access*, vol. 7, p. 65 811–65 820, 2019.
[27] Z. Zhang, Y. Gao, Y. Liu, and Z. Li, "Performance evaluation of shortened transmission time interval in LTE networks," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2018, pp. 1–5.
[28] 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. Conf. Netw. Commun. (EuCNC)*, Jun. 2019, pp. 597–602.
[29] 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)*, Sep. 2019, pp. 1–7.
[30] R. Nissel and M. Rupp, "OFDM and FBMC-OQAM in doubly-selective channels: Calculating the bit error probability," *IEEE Commun. Lett.*, vol. 21, no. 6, pp. 1297–1300, Jun. 2017.

[31] Y. Li and L. J. Cimini, "Bounds on the interchannel interference of OFDM in time-varying impairments," *IEEE Trans. Commun.*, vol. 49, no. 3, pp. 401–404, Mar. 2001.
[32] V. D. Nguyen and H.-P. Kuchenbecker, "Intercarrier and intersymbol interference analysis of OFDM systems on time-varying channels," in *Proc. 4th IEEE Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, 2003, pp. 140–144.
[33] M. Simko, S. Pendl, S. Schwarz, Q. Wang, J. C. Ikuno, and M. Rupp, "Optimal pilot symbol power allocation in LTE," in *Proc. IEEE Veh. Technol. Conf. (VTC Fall)*, Sep. 2011, pp. 1–5.
[34] L. Marijanovic, S. Schwarz, and M. Rupp, "Intercarrier interference of multiple access UFMC with flexible subcarrier spacings," in *Proc. 25th Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2017, pp. 888–892.
[35] *5G; NR; Base Station (BS) Radio Transmission and Reception*, document TS 38.104, 3rd Generation Partnership Project, 2018.
[36] F. Schaich and T. Wild, "Waveform contenders for 5G-OFDM vs. FBMC vs. UFMC," in *Proc. 6th Int. Symp. Commun., Control Signal Process. (ISCCSP)*, 2014, pp. 457–460.
[37] V. Vakilian, T. Wild, F. Schaich, S. ten Brink, and J.-F. Frigon, "Universal-filtered multi-carrier technique for wireless systems beyond LTE," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2013, pp. 223–228.
[38] X. Wang, T. Wild, F. Schaich, and S. ten Brink, "Pilot-aided channel estimation for universal filtered multi-carrier," in *Proc. IEEE 82nd Veh. Technol. Conf. (VTC-Fall)*, Sep. 2015, pp. 1–5.
[39] L. Zhang, P. Xiao, and A. Quddus, "Cyclic prefix-based universal filtered multicarrier system and performance analysis," *IEEE Signal Process. Lett.*, vol. 23, no. 9, pp. 1197–1201, Sep. 2016.
[40] J. Wen, J. Hua, W. Lu, Y. Zhang, and D. Wang, "Design of waveform shaping filter in the UFMC system," *IEEE Access*, vol. 6, pp. 32300–32309, 2018.
[41] S. Geng, X. Xiong, L. Cheng, X. Zhao, and B. Huang, "UFMC system performance analysis for discrete narrow-band private networks," in *Proc. IEEE 6th Int. Symp. Microw., Antenna, Propag., EMC Technol. (MAPE)*, Oct. 2015, pp. 303–307.
[42] *NR; Physical Channels and Modulation*, document TS 38.211, 3rd Generation Partnership Project, 2018.
[43] T. Pinto, C. Alves, R. Mansi, and J. V. de Carvalho, "Solving the multi-scenario max-min knapsack problem exactly with column generation and branch-and-bound," *Math. Problems Eng.*, vol. 2015, pp. 1–11, Feb. 2015.
[44] J. R. Tebboth, "A computational study of Dantzig–Wolfe decomposition," Ph.D. dissertation, Univ. Buckingham, Buckingham, U.K., 2001.
[45] F. Vanderbeck, "On Dantzig-Wolfe decomposition in integer programming and ways to perform branching in a branch-and-price algorithm," *Oper. Res.*, vol. 48, no. 1, pp. 111–128, Feb. 2000.
[46] *Study on Channel Model for Frequencies From 0.5 to 100 GHz*, document TR 38.901, 3rd Generation Partnership Project, Jan. 2018.
[47] S. Pratschner, B. Tahir, L. Marijanovic, M. Mussbah, K. Kirev, R. Nissel, S. Schwarz, and M. Rupp, "Versatile mobile communications simulation: The vienna 5G link level simulator," *EURASIP J. Wireless Commun. Netw.*, vol. 2018, no. 1, p. 2226, Dec. 2018.
[48] *5G;NR;User Equipment (UE) Radio Transmission and Reception;Part 1: Range 1 Standalone*, document TS 38.101, 3rd Generation Partnership Project, 2018.
[49] J. van den Brand, "A deterministic linear program solver in current matrix multiplication time," 2019, *arXiv:1910.11957*. [Online]. Available: http://arxiv.org/abs/1910.11957
[50] M. Simko, P. S. R. Diniz, Q. Wang, and M. Rupp, "Adaptive pilot-symbol patterns for MIMO OFDM systems," *IEEE Trans. Wireless Commun.*, vol. 12, no. 9, pp. 4705–4715, Sep. 2013.

LJILJANA MARIJANOVIĆ (Student Member, IEEE) received the M.Sc. degree in electrical and computer engineering from the Faculty of Technical Sciences, Novi Sad, Serbia, in 2014. She is currently pursuing the Ph.D. degree in telecommunications with TU Wien. Since August 2015, she has been employed as a Project Assistant with TU Wien. Her research interests include wireless communications, link level modeling and simulations, and signal processing.

STEFAN SCHWARZ (Senior Member, IEEE) received the Ph.D. degree in telecommunications engineering from the Technical University of Vienna (TU Wien), Austria, in 2013. He currently holds a Tenure Track position as an Assistant Professor with the Institute of Telecommunications, TU Wien, where he also heads the Christian Doppler Laboratory for Dependable Wireless Connectivity for the Society in Motion. His research interests include wireless communications, signal processing, and channel modeling. He is an Associate Editor of IEEE Access and *EURASIP JWCN*.

MARKUS RUPP (Fellow, IEEE) received the Dipl.Ing. degree from the University of Saarbrücken, Germany, in 1988, and the Dr.Ing. degree from Technische Universität Darmstadt, Germany, in 1993. Until 1995, he held a Postdoctoral position with the University of California at Santa Barbara, Santa Barbara, CA, USA. From 1995 to 2001, he was with the Wireless Technology Research Department, Bell Labs, Holmdel, NJ, USA. Since 2001, he has been a Full Professor of digital signal processing in mobile communications with TU Wien.