IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 4, 2020, accepted February 25, 2020, date of publication March 2, 2020, date of current version March 16, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2977773

A RAN Resource Slicing Mechanism for Multiplexing of eMBB and URLLC Services in OFDMA Based 5G Wireless Networks

PRAVEENKUMAR KORRAI, (Student Member, IEEE),
EVA LAGUNAS, (Senior Member, IEEE),
SHREE KRISHNA SHARMA, (Senior Member, IEEE),
ASHOK BANDI, (Student Member, IEEE),
SYMEON CHATZINOTAS, (Senior Member, IEEE),
AND BJÖRN OTTERSTEN, (Fellow, IEEE)
Interdisciplinary Center for Security, Reliability and Trust (SnT), University of Luxembourg, 4365 Esch-sur-Alzette, Luxembourg

Corresponding author: Praveenkumar Korrai (praveen.korrai@uni.lu)

This work was supported by the Luxembourg National Research Fund (FNR), under the AFR research grant entitled "Learning-Assisted Cross-Layer Optimization of Cognitive Communnication networks" and "5G-Sky: Interconnecting the Sky in 5G and Beyond – A Joint Communication and Control Approach."

ABSTRACT Enhanced mobile broadband (eMBB) and ultra-reliable and low-latency communications (URLLC) are the two main expected services in the next generation of wireless networks. Accommodation of these two services on the same wireless infrastructure leads to a challenging resource allocation problem due to their heterogeneous specifications. To address this problem, slicing has emerged as an architecture that enables a logical network with specific radio access functionality to each of the supported services on the same network infrastructure. The allocation of radio resources to each slice according to their requirements is a fundamental part of the network slicing that is usually executed at the radio access network (RAN). In this work, we formulate the RAN resource allocation problem as a sum-rate maximization problem subject to the orthogonality constraint (i.e., service isolation), latency-related constraint and minimum rate constraint while maintaining the reliability constraint with the incorporation of adaptive modulation and coding (AMC). However, the formulated problem is not mathematically tractable due to the presence of a step-wise function associated with the AMC and a binary assignment variable. Therefore, to solve the proposed optimization problem, first, we relax the mathematical intractability of AMC by using an approximation of the non-linear AMC achievable throughput, and next, the binary constraint is relaxed to a box constraint by using the penalized reformulation of the problem. The result of the above two-step procedure provides a close-to-optimal solution to the original optimization problem. Furthermore, to ease the complexity of the optimization-based scheduling algorithm, a low-complexity heuristic scheduling scheme is proposed for the efficient multiplexing of URLLC and eMBB services. Finally, the effectiveness of the proposed optimization and heuristic schemes is illustrated through extensive numerical simulations.

INDEX TERMS Network slicing, RAN radio resource allocation, sum-rate maximization, scheduling, eMBB and URLLC.

I. INTRODUCTION

The third and fourth generations (3G and 4G) of wireless networks have already revolutionized social behaviors through empowering the generalization of social networking on wireless mobile devices [2]. In order to further improve our cities, living environment, and industries, the next generation of wireless networks requires to support a large variety of services and applications with different requirements. Towards this achievement, the fifth-generation (5G) of wireless networks is expected to support the three major usage scenarios, which, according to ITU-R, are classified as enhanced mobile broadband (eMBB), ultra-reliable and low latency communications (URLLC) and massive machine-type communications [3]–[6]. A brief characterization of theses services is provided as follows: eMBB service requires higher data

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

rates to further improve the current mobile services such as high definition (HD) video and virtual reality (VR); URLLC service concentrates on supporting low-latency transmissions of small packets with high reliability and it covers applications such as autonomous vehicles, industrial automation, and vehicular communications; mMTC supports the services that connect a massive number of devices where each device transmits small data packets intermittently and it covers the applications like smart cities.

Out of the aforementioned three services, the main two services to be supported in the next release of 5G wireless networks are eMBB and URLLC [7], and thus are considered in this paper. However, the current one-size-fits-all network model is not suitable to accommodate these services [8]. Furthermore, accommodating these different wireless services in the same physical network while assuring their potential co-existence is also a major challenge. To address this problem, the next generation of wireless networks is expected to exploit the highly flexible and scalable network architectures to support the diverse applications from the aforementioned different services. In this context, recently, network slicing has emerged as a promising network architecture for allocating resources to different wireless services with diverse quality-of-service (QoS) needs [9]. In this approach, the common physical network infrastructure is sliced into multiple end-to-end logical networks, where each logical network acts as a dedicated network for a specific service. Specifically, each logical network or a network slice consists of a collection of particular radio access mechanisms and network functions and needs to be isolated from other slices that can be acquired through logical partitioning and radio resource virtualization.

Network slicing is executed both on the Radio Access Network (RAN) and the Core Network (CN). Thus, both parts of the network require to be sliced into multiple overlaid instances to serve the different types of services. In this paper, we focus on RAN slicing, whose challenges reside in the management of heterogeneous traffic demands coming from a variety of multiple users and services, and the limited available radio resources to satisfy these needs. To this end, the dynamic allocation of radio resources aligned with the instantaneous user traffic demands represents a major challenge. In this paper, we address the RAN slicing problem by dynamically assigning radio resource blocks (RBs) to each user according to its traffic demand in the network. In the following sub-sections, we review the related works from the literature and highlight the contributions of this paper.

A. REVIEW OF RELATED WORKS

Dynamic RAN radio resource allocation mechanisms have received recently increasing attention in the literature. Some of the relevant research studies are summarized in the following. In order to provide services to different service providers, a resource sharing approach for an efficient slicing of LTE network into multiple virtual networks (VNs) has been studied in [11]. A new slicing and scheduling technique has been introduced in [12] for wireless virtual networks (WVNs), which dynamically assign a specific number of RBs to each VN to assure services to its users. The authors of [13] have studied the problems of admission control and resource provisioning in Orthogonal Frequency Division Multiple Access (OFDMA) based WVNs. In [14], the energy-efficient sub-carrier and power allocation strategy with WVN has been proposed for a single-cell OFDMA system. Though the above-mentioned works have considered some constraints to maintain isolation between the slices, they may not be applicable for the next generation networks due to lack of considerations for the end-to-end QoS requirements of different services.

As aforementioned, the 5G NR has to support multiple number of services and a huge variety of applications. Using the simplified queuing analysis, in [15], the authors have shown that the dynamic multiplexing of URLLC and eMBB traffic significantly improves the total resource efficiency of a wireless system. The authors of [16] have investigated the radio-channel and QoS aware packet scheduling technique for the multiplexing of radio eMBB and URLLC services on the same radio resources in accordance with their demanding QoS requirements. In this work [16], the proposed scheduling algorithm dynamically adjusts the BLEP of URLLC transmissions according to the available traffic load and also a new channel quality indicator (CQI) estimation mechanism was introduced to improve the accuracy of the URLLC link adaption process. Furthermore, to advance the scheduler given in [15], a packet-size and control channel aware resource allocation method has been investigated in [17]. However, the proposed heuristic scheduling algorithms in [16], [17] satisfy the QoS requirements of URLLC service through prioritizing the service (i.e., allocation of large number of resources), but cannot guarantee the isolation between the service slices under the high URLLC loads (i.e., eMBB users may not get enough RBs under the high URLLC traffic).

On the another hand, puncturing based schemes have been recently proposed in the literature to eliminate the queuing delay of randomly occurred URLLC traffic through placing the URLLC traffic on the ongoing eMBB traffic [6], [18]–[20]. In this regard, [6] uses information theoretic results to achieve expressions for the average eMBB rates under URLLC puncturing for different decoding methods for uplink eMBB traffic superposed or punctured by the URLLC users. However, the authors of [6] have not considered the design of joint scheduling schemes for eMBB and URLLC data traffic. In [18], a punctured scheduling mechanism has been introduced for the transmission of latency critical traffic on a shared channel with the eMBB traffic, wherein the authors have utilized recovery techniques for eMBB transmissions, a service-specific heuristic scheduling algorithm and link adaption to increase the efficiency of the proposed scheme. The authors of [19] have proposed an online joint scheduling algorithm to improve the network utility of eMBB while assuring the stringent requirements of URLLC and studied optimal online joint URLLC/eMBB schedulers within the broad class of channel state dependent but mini-slot-homogeneous policies by using a more general class of convex and threshold loss models.

Furthermore, a risk-sensitive approach for the efficient allocation of radio resources to the eMBB and URLLC transmissions has been investigated in [20]. This work formulated the resource allocation problem as an optimization problem to maximize the overall eMBB data rate while assuming the risk of eMBB using the conditional value at risk (CVaR) function as a risk measure. In general, the punctured scheduling algorithms proposed in [6], [18]–[20] prioritize the URLLC service when the URLLC traffic arrives sporadically and places the URLLC traffic on ongoing eMBB traffic that leads to the isolation problems and also greatly reduces the eMBB data rate and reliability at the higher URLLC traffic. Also, the aforementioned puncturing mechanisms decrease the decoding ability due to the potential inter-user interference and also increase the control channel (CCH) overhead due to the utilization of extra dedicated CCHs for indicating the URLLC overlapping positions to the eMBB receiver. Moreover, most of the aforementioned works [15]–[20] have proposed the heuristic scheduling algorithms for dynamic multiplexing of eMBB and URLLC users and also considered the mini-slot (i.e., 2 OFDM symbols) based scheduling process. Nevertheless, none of works have targeted a design of the RAN resource slicing mechanism that efficiently optimize the currently available LTE standard radio resources (i.e., 0.5ms each transmission time interval (TTI), 1ms sub-frame and 10ms frame) between eMBB and URLLC services according to their isolation constraints and QoS requirements such as latency, reliability and minimum data rate. Moreover, most of the works e.g. [15], [19], [20] did not consider the AMC based link adaption process.

B. CONTRIBUTIONS

In the above context, different from the existing works, in this work, we propose a RAN resource slicing technique for the efficient multiplexing of eMBB and URLLC services in wireless networks by considering an AMC scheme. Subsequently, this resource slicing problem is formulated as an optimization problem to maximize the sum rate of all users, while satisfying the isolation constraint and stringent QoS constraints of the users such as latency and reliability. In this AMC based resource optimization problem, each RB's achievable data rate is measured by the chosen modulation and coding scheme (MCS) instead of Shannon rate formula. Furthermore, different signal-to-noise ratio (SNR) levels are modeled based on the specific user channel conditions to choose the MCS in accordance to their target block error rate (BLER) which depends on the reliability constraint of each service. This work builds on the author's previous publication [1]. In [1], a simple network model (without consideration of queues and traffic models) is considered for the analysis of slicing based resource allocation while this paper considers a queue based system model and different traffic models for eMBB and URLLC. Furthermore, in [1], a hard threshold-based approach is considered to relax the binary constraint in the optimization problem while a penalized formulation is considered in this work for the mathematical tractability. Besides, we propose a heuristic algorithm to solve the formulated optimization problem, which was missing in [1]. The main contributions of this paper are summarized in the following.

- Firstly, an AMC based resource allocation scheme is proposed for the dynamic multiplexing of eMBB and URLLC users on the same RAN infrastructure of a wireless network. In this model, the data traffic of each eMBB user is considered to be full-buffered with an infinite packet size and the data traffic of each URLLC user is generated using the 3GPP's FTP3 model [3] with B bytes of packet size.
- Secondly, we formulate the above resource allocation problem as an optimization problem to maximize the overall sum-rate of the network while satisfying the latency-related constraint of URLLC users and the minimum rate constraint of the eMBB users. However, due to the presence of the AMC scheme and binary assignment variable, the formulated optimization problem becomes analytically intractable. To address this issue, as a first step, by considering the two approximation functions, the step-wise optimization problem is transformed into a continuous linear problem. In the next step, a penalized formulation is considered to relax the binary constraint (i.e., assignment variable). Using the above two steps, we provide the solution to the proposed optimization problem for RAN slicing.
- Thirdly, a heuristic scheduling algorithm is proposed to overcome the complexity, time-consumption and in-feasibility problems of the proposed optimization problem.
- Finally, the performances of the proposed schemes are evaluated and compared through extensive simulations to illustrate their capability to perform a dynamic allocation of resources for different services while satisfying their reliability, latency and minimum rate requirements.

The remainder of the paper is organized as follows. Next section provides the system model and a detailed description of the radio resources considered in the scheduling process. In Section III, the proposed RAN resource optimization problem and its solution are presented. A low-complexity heuristic algorithm for the dynamic resource allocation is presented in Section IV. Section V provides the performance analysis of the proposed scheduling algorithms using the extensive numerical evaluations. Finally, the conclusions are drawn in Section VI.

II. SYSTEM MODEL

We consider the downlink (DL) OFDMA scenario of a single-cell cellular network, where a base station (BS) is located at the center of the cell, and M user equipment (UEs) are distributed randomly across the network area as shown

TABLE 1. Summary of notations.

Symbol	Definition
$\mathcal{U}$	Set of available UEs
M	Total number of UEs in the network
$\mathcal{U}_1$	Set of available eMBB users
L	Total number of eMBB UEs in the network
$\mathcal{U}_2$	Set of available URLLC users
K	Total number of URLLC UEs in the network
λ	URLLC packets arrival rate (packets per second)
B	URLLC packet size (in bytes)
$\mathcal{B}_{RB}$	The available bandwidth for each RB
T	Total available frames for scheduling
l	Set of available frames $\{1,2..,T\}$
N	Number of total TTIs in a frame
t	Set of TTIs in a frame $\{1,2..,N\}$
F	Number of available sub-bands (or RBs) per TTI
ρ	Number of available REs per RB (after accounting for RSs)
$\mathcal{R}^u_{mbb}$	Bit-rate of u^{th} eMBB user
$\mathcal{R}^u_{llc}$	Bit-rate of u^{th} URLLC user
$\mathbb{R}_{tot}$	Total sum-rate of the network
w_u	Weight on u^{th} URLLC user
$R^u_{t,f}$	Bit-rate of u^{th} user on sub-band 'f'at TTI 't
$\gamma^u_{t,f}$	Received SNR of u^{th} user in sub-band 'f'at TTI 't
$\mathcal{M}_{mcs}$	Number of available distinct MCSs
P_{max}	Available total power at BS
P	Allocated power to each RB
$h^u_{t,f}$	Channel gain of u^{th} user on a sub-band 'f'at TTI 't'
$d_{BS,u}$	distance between the BS and UE
σ^2	Noise Power
R_{min}	Minimum number bits per RB for URLLC
φ	SE of the selected MCS
$Q^{(l)}_u$	Queue length of u^{th} user on the l^{th} frame
R_{th}	Rate threshold for eMBB users
β_{mbb}	BLER target for eMBB users
β_{llc}	BLER target for URLLC users
Γ_{mbb}	SNR gap for eMBB users
Γ_{llc}	SNR gap for URLLC users
ζ_1	Penalty parameter
$\mathcal{P}(\cdot)$	Penalty function
$\nabla\mathcal{P}(\cdot)$	Gradient of $\mathcal{P}(\cdot)$

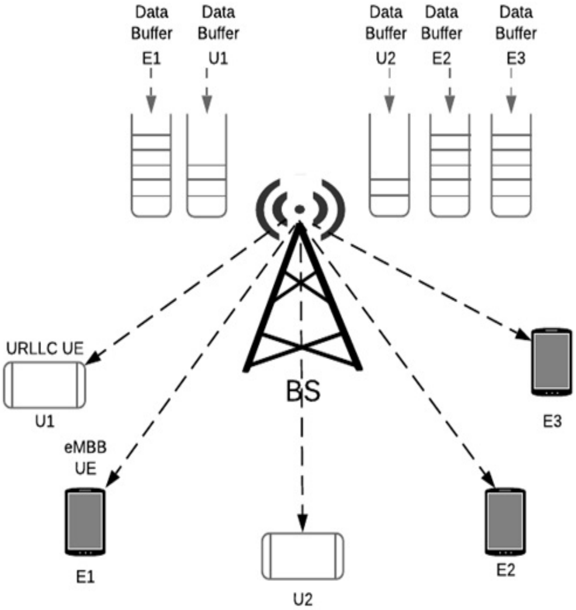


FIGURE 1. Illustration of a DL single-cell cellular network serving heterogeneous services (URLLC and eMBB).

in Fig. 1. The distributed UEs are associated with different types of services such as eMBB and URLLC (i.e., different services exist in the network). In the network, the UE associated with the URLLC service generates bursts of small packets of B bytes following the Poisson Point Process (PPP) with the arrival rate of λ [packets/sec]. This traffic model is termed as FTP3 in 3GPP [4]. Furthermore, the UE associated with the eMBB service generates continuous traffic (i.e., full-buffer traffic) with the infinite packet size. Traffic requested by the users located within the considered RAN undergo the specific admission and congestion control schemes implemented in the L3 and above OSI layers. These schemes allow the base station to make independent decisions about what set of packets to accept and/or what bit rates to offer to elastic traffic users competing for the same bandwidth. In general, if traffic loads are manageable, these are queued and scheduled in due time according to each slice QoS and priority. In our case, and as the general assumption in the literature, e.g. [15], we assume that the data from higher layers are received at the serving BS and stored in their respective user-specific transmission buffer until they get to be served as shown in Fig. 1. At each TTI, the proposed technique analyses the queues' status and provides the allocation of RBs according to the formulated mathematical optimization problem. However, the buffer congestion may show a significant effect on the QoS requirements. To avoid these disadvantages, we make the following assumptions: (1) Congestion control mechanisms are applied at higher layers that detect potential congestion and temporally reduce the transmission data rate, (2) infinite buffer size that avoids the packet loss due to buffer overflow. Note that the URLLC packet latency measurements provided in the manuscript correspond to the gap (measured in TTIs) between the time that the particular URLLC packet has entered the queue and the time that the packet has been scheduled and left the queue. Furthermore, these available packets in the buffer (i.e., queue) of each UE are served by the BS on the First-In and First-Out (FIFO) basis.

The BS serves all the UEs in the cell, indexed by $\mathcal{U} = \{1, 2, \ldots .M\}$, through the available radio resources. In our analysis, the available radio resources are two-dimensional (2D), i.e., including both time and frequency domains.The DL frequency bandwidth is partitioned into F sub-bands indexed by $f = \{1, 2, \ldots F\}$ and the time dimension is divided into transmission time intervals (TTIs) indexed by $t = \{1, 2, \ldots N\}$ with the duration of 0.5 ms as shown in Fig. 2. Thus, a total F number of RBs are available for DL transmission in a one-time slot. As defined in 3GPP 5G-NR [10], an RB is the minimum time-frequency resource that can be allocated to a specific user, which consists of 7 OFDM symbols and 12 consecutive sub-carriers (for a complete bandwidth of 180 KHz) [21], [22]. Therefore, each RB (i.e., 12 sub-carriers, 7 OFDM symbols) includes 84 Resource Elements (REs), after accounting for the reference signals overhead, approximately $\rho = 60$ REs per RB

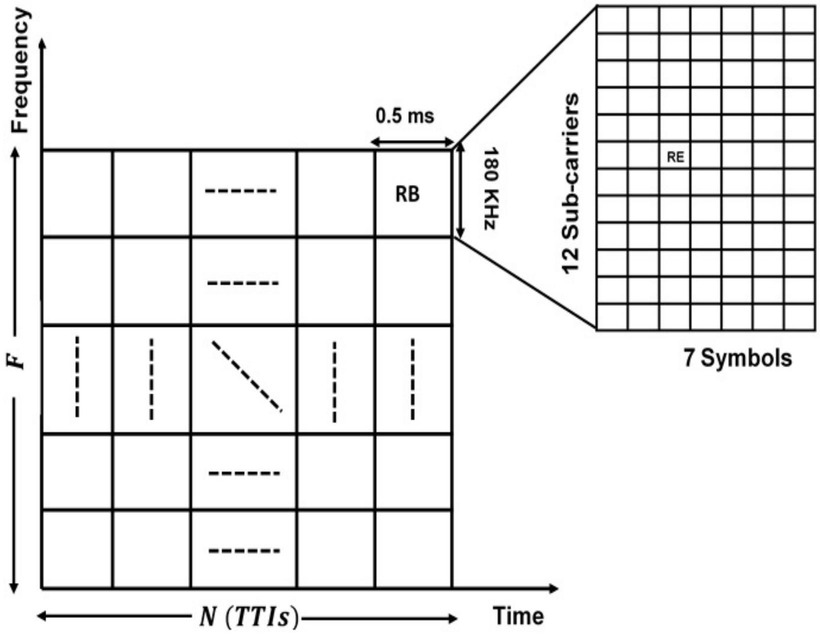


FIGURE 2. Illustration of time and frequency radio resource grid in the considered system scenario.

TABLE 2. Modulation and Coding Schemes (MCS) for eMBB and URLLC services with different BLERs.

MCS	Modulation	Code Rate	SNR Threshold [dB] BLER 0.1	SNR Threshold [dB] BLER 0.001	Efficiency [bits/Symbol]
MCS1	QPSK	1/12	-6.5	-2.5	0.15
MCS2	QPSK	1/9	-4.0	0.0	0.23
MCS3	QPSK	1/6	-2.6	1.4	0.38
MCS4	QPSK	1/3	-1.0	3.0	0.60
MCS5	QPSK	1/2	1.0	5.0	0.88
MCS6	QPSK	3/5	3.0	7.0	1.18
MCS7	16QAM	1/3	6.6	10.6	1.48
MCS8	16QAM	1/2	10.0	14	1.91
MCS9	16QAM	3/5	11.4	15.4	2.41
MCS10	64QAM	1/2	11.8	15.8	2.73
MCS11	64QAM	1/2	13.0	17	3.32
MCS12	64QAM	3/5	13.8	17.8	3.90
MCS13	64QAM	3/4	15.6	19.6	4.52
MCS14	64QAM	5/6	16.8	20.8	5.12
MCS15	64QAM	11/12	17.6	21.6	5.55

are available, where each RE comprises of a sub-carrier and an OFDM symbol [23], [24]. In this paper, we consider the 4G LTE standard radio-frame numerology[1] for radio resource allocation to DL transmissions, which is also considered as a candidate numerology for upcoming 5G systems. First, each user computes the channel quality indicators (CQIs) for all the available RBs and feeds back its CQIs to the BS. If the RB is allocated to the specific UE, the AMC method allows the wireless system to choose the appropriate MCS according to the received CQI. Based on the BLER or probability of error, and the received CQI feedback from the user, the minimum SNR threshold is set to obtain the appropriate MCS. For instance, if MCS14 is chosen, the SE of the MCS14 is 5.12 bits/symbol and each RE carries 5.12 bits (using MCS Table. 2). As a result, each RB carries $60 \times 5.12 = 307.2$ bits on average.

III. PROBLEM FORMULATION AND PROPOSED SOLUTION

The fifth-generation (5G) wireless networks are expected to support users from multiple services. These services are mainly classified as latency-critical and rate-based services. The latency-critical service needs to satisfy delay constraint, while the rate-based service requires a minimum rate to support continuous traffic demand. Therefore, an efficient scheduling mechanism is essential to allocate the resources for these two type of services by satisfying their requirements (i.e., aforementioned constraints). In other words, in this work, we develop a slice-aware RAN radio resource

[1] A mixture of different numerology defined in [2] is out of scope and left for the future work.

allocation mechanism [30], which shares the available radio resources by considering the specific constraints for eMBB and URLLC users to make sure that the performance guarantees of slices are satisfied and the slices do not adversely affect the each other's performance. Particularly, in this section, we address the AMC based sum-rate maximization problem for the dynamic allocation of radio resources to the wireless system consisting of multiple services.

A total T number of frames are considered for the scheduling process, wherein each frame consists of F number of RBs and N number of TTIs. The problem consists in assigning the total number of $N \times F$ RBs to the active users. During the scheduling round (i.e., for every TTI), each RB is assigned to a single user. Denoting $x^u_{t,f}$ a binary assignment variable, which is 1 if the RB (t, f) is allocated to the user u, otherwise, it is 0. Then, the binary constraint is mathematically written as

$$x^u_{t,f} = \begin{Bmatrix} 1; & \text{If RB } (t,f) \text{ is allocated to user '}u\text{'} \\ 0; & \text{Otherwise} \end{Bmatrix} \tag{C1}$$

In this work, we assume the presence of both the eMBB and URLLC users in the network. The sets $\mathcal{U}_1 = \{1, 2, \ldots, L\}$ and $\mathcal{U}_2 = \{1, 2, \ldots, K\}$, represent the sets of users associated with eMBB and URLLC services, respectively. The objective function, i.e., the total sum-rate of the network, is then given by

$$\mathbb{R}_{tot} = \sum_{u\in\mathcal{U}_1} \mathcal{R}^u_{mbb} + \sum_{u\in\mathcal{U}_2} w_u \mathcal{R}^u_{llc} \tag{1}$$

where $\mathcal{R}^u_{mbb}$, $\mathcal{R}^u_{llc}$ are the bit rates of the u^{th} eMBB and URLLC users, and w_u is the weight factor of the u^{th} URLLC user. In order to prioritize the URLLC user that accumulates more packets in its queue, w_u is included in the URLLC individual user rates. Now, w_u is expressed as

$$w_u = \frac{Q^{(l)}_u}{\sum\limits_{u\in U_2} Q^{(l)}_u} \tag{2}$$

where $Q^{(l)}_u$ represents the queue length of u^{th} user on the l^{th} frame measured in bits. A high w_u value indicates a high priority URLLC user. The bit rate of each user that belongs to either eMBB or URLLC service, $\mathcal{R}^u_s$, $s \in \{mbb, llc\}$ is computed as

$$\mathcal{R}^u_s = \sum_{t=1}^{N}\sum_{f=1}^{F} x^u_{t,f} R^u_{t,f}, \quad \begin{matrix} \forall\, u \in \mathcal{U}_1 : s = mbb; \\ \forall\, u \in \mathcal{U}_2 : s = llc; \end{matrix} \tag{3}$$

where the bit rate of user u operating in sub-band f at TTI t can be expressed as

$$R^u_{t,f} = \mathcal{B}_{RB} \cdot T \cdot \mathcal{F}(\gamma^u_{t,f}) \qquad \text{[bits]} \tag{4}$$

where $\mathcal{B}_{RB}$ is the bandwidth of an RB, T is the transmission time length of each slot and $\mathcal{F}(.)$ is the spectral efficiency (SE) of the selected MCS from Table. 2 according to the achievable SNR. In our study, $\mathcal{M}_{mcs}$ distinct MCSs are considered and the corresponding SNR levels of MCS for different BLER targets are provided in Table 2 [25]. Note that in Table 2, we have provided two different MCS depending on the BLER target. In particular, $BLER = 10^{-3}$ is used for to the URLLC service, while $BLER = 10^{-1}$ is used for the eMBB service. Also, perfect channel information is assumed at the BS for all users, and the total available power P_{max} is considered to be equally distributed over all available RBs for a TTI (i.e., allocated power to each RB is $P = P_{max}/F$). Then, the received SNR ($\boldsymbol{\gamma}^u_{t,f}$) of the u^{th} user in sub-band f at TTI t can be expressed as

$$\boldsymbol{\gamma}^u_{t,f} = \frac{P|h^u_{t,f}|^2 d^{-\alpha}_{BS,u}}{\sigma^2}, \tag{5}$$

where $h^u_{t,f}$ represents the channel gain of user u on a sub-band f at the TTI t, $d_{BS,u}$ is the distance between the BS and the UE, α is the path loss exponent and σ^2 is the noise power. Note that the inter-cell interference is assumed to be mitigated using suitable interference avoidance mechanisms such as in [34], [35].

We maximize the above objective function given in (1) subject to the constraints as follows. The aforementioned binary constraint (C1) or decision variable maintains the allocation of RBs to users and the constraint (C2) assures that an RB is only allocated to a single user (i.e., called as the orthogonality constraint).

$$\sum_{u\in\mathcal{U}_1}\sum_{u\in\mathcal{U}_2} x^u_{t,f} \leq 1; \quad \forall\, t, f \tag{C2}$$

The next constraint (C3) introduces to control the transmission latency requirement of URLLC users.

$$\sum_{f=1}^{F}\sum_{t=kp+1}^{kp+p} x^u_{t,f} \geq 1; k = 0, 1, 2, 3.., k_{max}; \quad \forall u \in \mathcal{U}_2 \tag{C3}$$

More precisely, when a URLLC user is scheduled (i.e., $u \in \mathcal{U}_2$), constraint (C3) enforces that at least one RB for every p TTIs is scheduled for each URLLC users until the required number of TTIs to vacate the queue. When the number of available URLLC packets in the queue becomes very high, the scheduler needs to allocate RBs to the user by following the given constraint (C3) in order to vacate as much as possible number of packets. Otherwise, when the available packets in the queue are fewer, then the scheduler needs to assign RBs until the queue becomes empty. To do so, the variable k_{max} is defined as

$$k_{max} = \min\left(\left\lceil \frac{N}{p} \right\rceil - 1, \left\lceil \frac{Q^{(l)}_u}{\rho \cdot \varphi} \right\rceil - 1\right) \tag{6}$$

where φ represents the SE of the MCS that ensures that the whole packet can be sent. For example, assume that the queue length of the u^{th} URLLC user is 1024, and MCS13 is selected for the scheduled RB. Using the formula in (6), the variable k_{max} is computed as $\lceil \frac{1024}{60\times4.52} \rceil - 1 = 3$, and the required number of RBs for the assumed queue length is estimated as $\psi_{RB} = k_{max} + 1$, i.e., $3 + 1 = 4$. Assuming $p = 2$, a total $(p \times \psi_{RB})$ TTIs, i.e., $2 \times 4 = 8$ are required to vacate the queue of this particular URLLC user. Note that the number of scheduling TTIs should not exceed the available TTIs in a frame such that by setting the minimum condition as given in (6), we can assure the feasible scheduling process.

Furthermore, each of the assigned RB to the URLLC user should at least transmits a complete data packet. In this regard, constraint (C4) enforces the condition that every scheduled RB for the URLLC user transmits more than the minimum number of bits (i.e., one packet size denotes by R_{min} measured in bits). It means that when an RB is scheduled to a u^{th} URLLC user (i.e., $x^u_{t,f} = 1$), the resulting rate from that assigned RB should be greater than or equal to R_{min} number of bits. Moreover, we assume that the packet size of all URLLC users is same and the packet segmentation is not allowed. Thus, this constraint helps to transmit at least a packet through the assigned RB.

$$x^u_{t,f} \cdot R^u_{t,f} \geq x^u_{t,f} \cdot R_{min};\ u \in \mathcal{U}_2 \tag{C4}$$

For every URLLC user, the constraints (C3) and (C4) together ensures the transmission of at least one packet of data for every p time slots till to reach the k_{max} number of TTIs. Note that the backlogged packets in the queue are transmitted in the early next frame.

Finally, to ensure a minimum throughput for the eMBB service, the constraint (C5) confirms that every scheduled eMBB user at least transmit R_{th} number of bits for every frame.

$$\sum_{f=1}^{F}\sum_{t=1}^{N} x^u_{t,f} R^u_{t,f} \geq R_{th};\ u \in \mathcal{U}_1 \tag{C5}$$

The major objective of the proposed optimization problem is to maximize the total sum-rate of users associated with eMBB and URLLC services through performing dynamic RBs allocation subject to a set of constraints. After the formulation of max sum-rate, the term of sum-data rate for URLLC service is required in conjunction with the latency constraint in order to adhere to the URLLC requirements, which has to be implemented by taking into account the available queue lengths of URLLC users (and updating the weight at each scheduling time).Therefore, the optimization problem is formulated as a scheduling of the time-frequency radio resources to the different users according to their available traffics (i.e., queue lengths) and respective QoS requirements. Mathematically, the optimization problem is expressed as

$$P1: \quad \max_{\{x^u_{t,f}\}} \ \mathbb{R}_{tot} \tag{7}$$

$$\text{subject to } x^u_{t,f} \in \{0,1\}; \quad \forall u, t, f \tag{C1}$$

$$\sum_{u\in\mathcal{U}_1}\sum_{u\in\mathcal{U}_2} x^u_{t,f} \leq 1; \quad \forall\, t, f \tag{C2}$$

$$\sum_{f=1}^{F}\sum_{t=kp+1}^{kp+p} x^u_{t,f} \geq 1;$$

$$k = 0, 1, 2, 3.., k_{max};\ u \in \mathcal{U}_2 \tag{C3}$$

$$x^u_{t,f} \cdot R^u_{t,f} \geq x^u_{t,f} \cdot R_{min};\ u \in \mathcal{U}_2 \tag{C4}$$

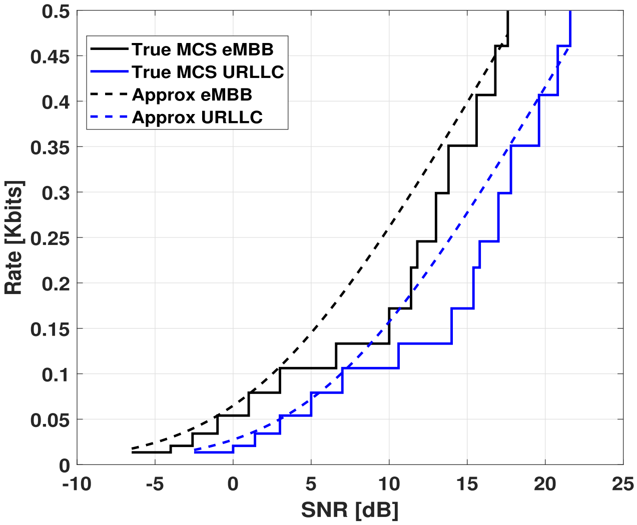


FIGURE 3. Rate of eMBB, URLLC services using true MCS given in Table 2 and approximated rate functions.

$$\sum_{f=1}^{F}\sum_{t=1}^{N} x^u_{t,f} R^u_{t,f} \geq R_{th};\ u \in \mathcal{U}_1 \tag{C5}$$

However, the presented function $\mathcal{F}(.)$ in (7), given by (4), is a step-wise function that makes the optimization problem mathematically intractable and complex to solve. Inspired by [26], to simplify the problem, using the received SNR and target BLER, we make use of two SE approximation functions (i.e., differentiable continuous functions) for eMBB and URLLC services that can be expressed as

$$\mathcal{F}_{mbb}(\boldsymbol{\gamma}^u_{t,f}, \beta_{mbb}) = \log_2(1 + \frac{\boldsymbol{\gamma}^u_{t,f}}{\Gamma_{mbb}}), \tag{8}$$

$$\mathcal{F}_{llc}(\gamma^u_{t,f}, \beta_{llc}) = \log_2(1 + \frac{\gamma^u_{t,f}}{\Gamma_{llc}}) \tag{9}$$

where $\Gamma_{mbb} = \frac{-\ln(5\beta_{mmb})}{0.45}$ and $\Gamma_{llc} = \frac{-\ln(5\beta_{llc})}{1.25}$ represent the SNR gaps, and β_{mbb} and β_{llc} represent the target BLER for eMBB and URLLC users, respectively. The proposed approximate functions are compared to the values achieved with AMC Table. 2 in Fig. 3, where we can observe that (8) and (9) provide a good approximation to the AMC step-wise functions.

The achievable data on each RB using the aforementioned approximation functions can be written as

$$r^{u,s}_{t,f} = B \cdot T \cdot \ \mathcal{F}_s(\gamma, \beta_s),\ s \in \{mbb, llc\}[bits] \tag{10}$$

Using (10), the bit-rate of each user that belongs to either eMBB or URLLC service can be reformulated as

$$\hat{\mathcal{R}}^u_s = \sum_{t=1}^{N}\sum_{f=1}^{F} x^u_{t,f} r^{u,s}_{t,f} \tag{11}$$

Now, using (11), the objective function sum-rate of all users can be reformulated as

$$\hat{\mathbb{R}}_{tot} = \sum_{u\in\mathcal{U}_1} \hat{\mathcal{R}}^u_{mbb} + \sum_{u\in\mathcal{U}_2} w_u \hat{\mathcal{R}}^u_{llc} \tag{12}$$

The optimization problem (P1) is now reformulated as

$$P2: \quad \max_{\{x^u_{t,f}\}} \quad \hat{\mathbb{R}}_{tot} \tag{13}$$

$$\text{subject to} \quad (C1), (C2), (C3), (C4) \text{ and } (C5) \text{ in } (P1) \tag{14}$$

Due to the binary constraint (C1), the problem P2 is combinatorial. Therefore, to avoid the combinatorial nature of $P1$, the assignment variable $x^u_{t,f}$ is relaxed to a box constraint between 0 and 1, and the relaxation penalty is added using the function $\mathcal{P}(x^u_{t,f})$ so that the relaxed problem produces the output favorable to either 0 or 1. The problem $P2$ is reformulated with the penalty parameter ζ_1 as

$$P3: \quad \max_{\{x^u_{t,f}\}} \left[\hat{\mathbb{R}}_{tot} + \sum_{u\in\mathcal{U}_1} \sum_{u\in\mathcal{U}_2} \sum_{f=1}^{F} \sum_{t=1}^{N} \zeta_1 \mathcal{P}(x^u_{t,f}) \right] \tag{15}$$

$$\text{subject to } (C1:)\ 0 \le x^u_{t,f} \le 1 \tag{16}$$

$$(C2), (C3), (C4) \text{ and } (C5) \text{ in } (P1) \tag{17}$$

By assuming $X = x^u_{t,f}$, the penalty function is defined as $\mathcal{P}(X) = (X^2 - X)$, which is a convex function in the region of [0, 1]. The function $\mathcal{P}(X)$ produces no penalty at $X = 0$ or 1 and increases the penalty as X moves away from 0 or 1 with the maximum penalty at $X = 0.5$. For example, when $X = 0.5$, the total incurred penalty is $(0.5)^2 - (0.5) = -0.25$. Further, by selecting the penalty parameter ζ_1 appropriately, the binary nature of X can be accomplished. Note that the objective function in $P3$ i.e., $\hat{\mathbb{R}}_{tot} - \left(- \sum_{u\in\mathcal{U}_1} \sum_{u\in\mathcal{U}_2} \sum_{f=1}^{F} \sum_{t=1}^{N} \zeta_1 \mathcal{P}(x^u_{t,f}) \right)$, is a difference of convex and concave functions. Thus, the problem $P3$ belongs to the class of difference of convex (DC) programming [27]. In this regard, we utilize an iterative algorithm based on convex-concave procedure (CCP) to solve the DC problem in (14). CCP is a dynamic tool to estimate the stationary point of the DC problems. In this algorithm, the following two steps are performed iteratively till its converges: (i) Assume X^{k-1} is the estimate of X in the $(k-1)^{th}$ iteration. In the k^{th} iteration, the affine approximation around the estimate of X^{k-1} is utilized to replace the convex part of the objective (i.e., the penalty summation part in (15)). (ii) The update X^{k+1} is acquired by solving the following convex problem:

$$P4: \quad \max_{\{X\}} [\hat{\mathbb{R}}_{tot} + \zeta_1 \sum_{u\in\mathcal{U}_1} \sum_{u\in\mathcal{U}_2} \sum_{f=1}^{F} \sum_{t=1}^{N} (X - X^{k-1}) \nabla \mathcal{P}(X)] \tag{18}$$

$$\text{subject to} (C1)\ 0 \le X \le 1; \forall u, t, f \tag{19}$$

$$(C2), (C3), (C4) \text{ and } (C5) \text{ in } (P1) \tag{20}$$

The above algorithm is based on the CCP framework, thus, a feasible initial point is enough to converge the algorithm to a stationary point [28]. Note that the initial feasible point need not to be binary, but it should satisfy all the other constraints in the optimization problem. Therefore, we set the initial feasible point by solving the problem $P1$ (i.e., without penalty function), wherein the constraint (C1) is relaxed between 0 and 1. Now, the problem $P4$ is a convex problem that can be solved by the standard optimization software tools such as CVX [29]. The proposed solution optimizes the resource allocation across both frequency and time simultaneously, covering the complete frame. The final output of the problem shows in each scheduling round which users should be served on which RB.

A. HEURISTIC ALGORITHM FOR SCHEDULING OF EMBB AND URLLC USERS

In this section, we propose a low-complexity greedy heuristic algorithm for scheduling the eMBB and URLLC users efficiently. This algorithm is proposed to maximize the overall sum-rate of the users in the network while prioritizing the URLLC users. A total T number of frames are considered for the complete scheduling process, wherein each frame consists of F number of RBs and N number of TTIs (i.e., $N \times F$ number of RBs for the complete frame). The proposed heuristic algorithm is executed for the given number frames T. In particular, at each frame l, $l = 1, \ldots T$, the URLLC users are firstly scheduled based on the best RB according to their channel conditions, followed by the scheduling on the remaining RBs for the eMBB, again based on their channel conditions. After completion of the scheduling of a particular frame l, the queues of the URLLC UEs are updated with the new arrived packets and the unscheduled packets of the previous frame. Clearly, the proposed heuristic provide the priority for URLLC users in order to satisfy the latency-related requirement. Summarizing, the scheduling process at each frame 'l' is executed in the steps as follows:

Step 1. SNRs Computation: Compute the received SNRs of all users on all the available RBs using (5) in the first TTI of every frame. Note that within 1 frame, assume that the channel remains temporal invariant, while it may change from carrier to carrier (i.e., RB to RB). Therefore, the computation of SNRs on the first TTI is sufficient for the complete scheduling process of a frame.

Step 2. RBs Assignment to Users: The URLLC service is prioritized than the eMBB service, hence, the active URLLC users schedule first according to the process as follows: first, identify the highest SNR received RB for every active URLLC user using Step 1 and next, compute the MCS and SE of the selected RBs by comparing the SNR values with the ones in Table 2. Note that, in this work the packet segmentation is not allowed for the URLLC users, so that the user can select the MCS according to the packet size (i.e., the user can select the lower MCS instead of achieved higher MCS, which is sufficient to transmit one packet). Then, using the updated queue and the SE of the selected RB, the number of required RBs to transmit the available traffic is computed as

$$k_u = \max(\lceil N/p \rceil, \lceil Q_u^{(l)}/\rho \cdot \varphi \rceil), \quad \forall u \in \mathcal{U}_2 \tag{21}$$

Later, assign the selected RB for every assuming p TTIs until to meet the k_u number of RBs. After scheduling the URLLC users in every TTI, remove the assigned RBs from the scheduling and update the available RBs for eMBB users scheduling. Now, the eMBB users first identify the highest SNR received RBs using Step 1 and next, estimate the MCS and SEs of the selected RBs by comparing the SNR values with the provided Table 2, and finally the selected RBs are assigned to every user in every scheduling TTI until the end of frame.

Step 3: Computation of Delivered Data and Queue Update: After completion of the scheduling process for the complete frame (i.e., up to N number of TTIs), compute the sum-rate (i.e., delivered data rate) of every user using the assigned number of RBs and SE of the selected RB. Next, estimate the undelivered data traffic by subtracting the delivered data traffic from the available data traffic (i.e., generated or original data traffic). Finally, update the queues with the undelivered data traffic, that is scheduled in the early next frame. The same procedure continues until the end of the frames. The complete procedure of heuristic scheduling scheme is summarized in Algorithm 1.

Algorithm 1 Heuristic Algorithm for Scheduling of eMBB and URLLC Users

```
Inputs:
  • Number of TTIs: $N$
  • Number of RBs: $F$
  • Class of eMBB and URLLC users: $\mathcal{U}_1$ and $\mathcal{U}_2$
  • Total number of users: $Num = \mathcal{U}_1 \cup \mathcal{U}_2$
  • Number of frames: $T$
  • Gen. data for $u^{th}$ URLLC user on $l^{th}$ frame : $G_u^{(l)}$
Initialization: $l = 1$ and queue: $\zeta_u^{(0)} = 0$;
while $l \leq T$ do
    for $u = 1 : Num$ do
        for $f = 1 : F$ do
            Estimate the received SNRs of the user on all RBs using (5)
        end for
        if $u \in \mathcal{U}_2$ then
            Find the highest SNR received RB:
                $f^* = \arg\max_{u \in \mathcal{U}_2} \gamma_{t,f}^u$
            Get the MCS and SE of the highest SNR received RB using Table 2;
            Get the queue length: $Q_u^{(l)} = G_u^{(l)} + \zeta_u^{(l-1)}$
            Compute the required number of RBs:
                $k_u = \max(\lceil N/p \rceil, \lceil Q_u^{(l)}/\rho \cdot \varphi \rceil)$
            Assign the selected RBs to the user for every p TTIs until to reach the $k_u$ RBs;
            Remove the selected RBs from the scheduling process in that TTI;
        end if
        if $u \in \mathcal{U}_1$ then
            Find the highest SNR received RBs:
                $f^* = \arg\max_{u \in \mathcal{U}_1} \gamma_{t,f}^u$
            Get the MCS and SE of the RB;
            Assign the selected RBs to the user in every TTI;
        end if
    end for
    Compute the delivered data of URLLC users:
        $R_u^{(l)} = k_u \cdot \rho \cdot \varphi; \forall u \in \mathcal{U}_2$
    Queue status update of of URLLC user:
        $\zeta_u^{(l)} = Q_u^{(l)} - R_u^{(l)}; \forall u \in \mathcal{U}_2$
    Compute the delivered data of eMBB users:
        $R_u^{(l)} = \sum_{i=1}^{\eta_u} \rho \cdot \varphi_u, \forall u \in \mathcal{U}_1$
    $l = l + 1$;
end while
```

B. COMPLEXITY ANALYSIS

The computational complexity of the optimization problem depends on the complexities of the following two procedures provided in Section III: (i) Convex-Concave procedure (CCP) involved in the convex problem $P4$, and (ii) Initial feasible point selection procedure using the convex problem $P1$. The convex problem $P4$ has MNF decision variables and $2MNF+NF+K(k_{max}+1)+M$ linear constraints. Therefore, the computational complexity of $P4$ is $\mathcal{O}((MNF)^3(2MNF + NF+K(k_{max}+1)+M))$. Similarly, the convex problem $P1$ has MNF decision variables and $2MNF + NF + Kk_{max} + 2K + L$ linear constraints. Hence, the computational complexity of $P1$ is $\mathcal{O}((MNF)^3(2MNF + NF + Kk_{max} + 2K + L))$ [32].

The complexity of heuristic algorithm is majorly due to the one while iterations and executing two *argmax*s. The complexity for executing *argmax* is proportional to the number of elements being sorted. Therefore, the complexity of heuristic algorithm is $\mathcal{O}(L+F+K)$ for each RB and the total complexity is $\mathcal{O}(FL + F^2 + FK)$ [33]. From this complexity analysis, it is easy to observe that the computational complexity of the optimization method is significantly larger compared to the heuristic method as expected.

IV. NUMERICAL EVALUATIONS

In this section, we simulate and compare the performance of the proposed optimization based and heuristic scheduling algorithms for the allocation of resources to the existing eMBB and URLLC users in the downlink single-cell OFDMA scenario.

A. SIMULATION ENVIRONMENT

We mainly concentrate on the resource scheduling for a downlink wireless network, where a single BS is deployed at the center of the cell coverage area with the radius of 250m, and L eMBB users and K URLLC users are distributed randomly within the cell coverage area. In this model, the channel between the BS and the each user is considered as a Nakagami-m fading channel. Also, the path loss exponent (α) is set to 3 for all the communication links. Our simulations are executed for 10 frames, where each frame consists of 20 TTIs (i.e., a total time of 10ms) and 100 RBs for each TTI. Each RB comprises of 12 sub-carriers,

TABLE 3. Simulation parameters.

Parameter	Value
Cell radius	$250m$
Number of users (M)	$15, 20, 50$
Number of eMBB users (L)	10
Number of URLLC users (K)	$5, 10, 40$
BS Transmit power (P_{max})	$10, 20, 30, 40$dBm
path loss exponent (α)	3
Channel	Nakagami-m fading model
Number of TTIs (N)	20
Each TTI length (T)	0.5ms
Total length of time-frame	10ms
RBs per TTI (F)	100
Number of sub-carriers per RB	12
Number of OFDM symbols per RB	7
Number of REs per RB	84
Reference signals overhead per RB	24
Each sub-carrier length	15KHz
Each RB's bandwidth	180KHz
Carrier Bandwidth	20 MHz
Traffic model for URLLC	FTP3 model
URLLC packet size (B)	32 bytes
Traffic model for eMBB	Full-buffered
eMBB packet size	Infinite

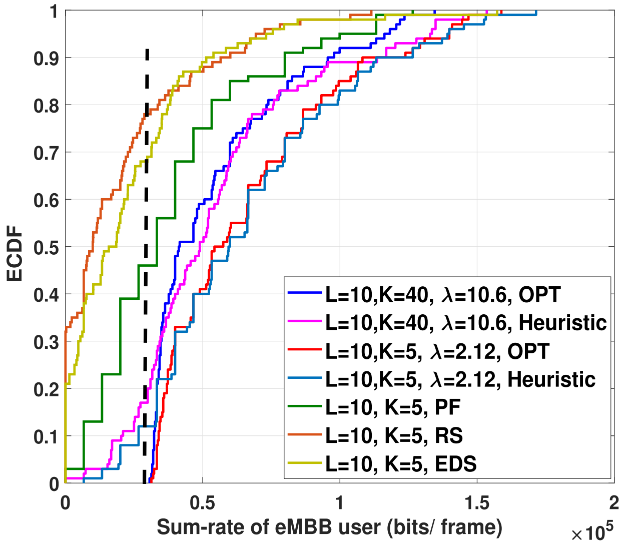


FIGURE 4. Cumulative distribution of achieved eMBB rates (per frame) for different URLLC loads using the proposed scheduling algorithms and baseline methods.

7 OFDM symbols, and a total of 60 REs after accounting for the reference signals. Moreover, each sub-carrier has a carrier-spacing of 15 KHz. Hence, the bandwidth of each RB is 180 KHz and the available complete bandwidth for the BS is 20 MHz. Also, we assume that the additive white Gaussian noise power on each sub-band is 10^{-10}W. Importantly, to satisfy the reliability of each service, we assume a high BLER target (i.e., $\beta_{mbb} = 10^{-1}$) for eMBB users as compared to URLLC user's BLER target (i.e., $\beta_{llc} = 10^{-3}$). Further, we consider that each UE has a buffer to store the generated packets prior to serve. In this network model, we assume that each URLLC UE generates the small bursts of data following the FTP3 model with the mean arrival rate of λ payloads per frame (i.e., for example $\lambda = 2.12$ packets/10ms, equal to 212 packets/sec), and each eMBB user generates the packet with an infinite size. The complete set of simulation parameters is provided in Table 3.

B. RESULTS AND DISCUSSIONS

We compare the performance of the proposed methods in Section III with the performance of baseline methods random scheduler (RS) (i.e., distributes RBs randomly to URLLC and eMBB users), equally distributed scheduler (EDS) (i.e., distributes RBs equally to URLLC and eMBB users) [31] and proportional fair (PF) scheduler [15] in terms of achieved delivered data rate for eMBB users. In Fig. 4, we illustrate the empirical cumulative distribution function (ECDF) of achieved delivered data rates of eMBB users on every frame using the proposed methods, RS, EDS, and PF. From the results in Fig. 4, it can be observed that the baseline methods RS, EDS, and PF fail to achieve the minimum delivered rate (per frame) for many of eMBB users. Also, it is evident from the results that using the heuristic scheduling algorithm, some of the eMBB users fail to satisfy the constraint (C5) (i.e., the minimum rate requirement, $R_{th} = 30$ Kbits per frame). The heuristic algorithm first greedily assigns RBs to URLLC users by giving them priority, and subsequently assigns the remaining RBs (i.e., left after scheduling of the URLLC users) to the eMBB users, which are not enough sometimes to achieve the minimum rate. Also, in the eMBB users scheduling process, the heuristic algorithm greedily assigns the RBs to the users which receive the highest SNRs, so that the users with lowest SNRs cannot get the required number of RBs to achieve the minimum rate. In contrast, the optimization based scheduling algorithm provides the RBs to users by respecting the isolation between service slices, so that it satisfies the minimum rate condition of eMBB users regardless of the channel conditions. The results confirm that the heuristic algorithm, although lighter in complexity compared to the proposed optimization procedure, it fails in satisfying isolation and minimum rate requirements.

In Fig. 5, we evaluate the ECDF of the latency in the delivered URLLC packets measured as the gap between the TTI that the packets have entered the queue and the TTI the packet has been scheduled and left the queue. Therefore, the total packet latency time is computed as the sum of packet waiting time in the queue, the required time to assign the RB and data transmission (i.e., scheduling time). From the results in Fig. 5, it is observed that adjusting the value of p in (C3) (i.e., difference between the scheduling time intervals is less) the total latency can be reduced. For instance, by assigning an RB to URLLC user for every 2 TTIs, 40% of packets experience approximately 3 TTIs less latency as compared to the assignment of RB to URLLC users for every 4 TTIs. As expected, the heuristic scheduling algorithm shows a slightly better performance compared to the optimization based scheduling algorithm in terms of latency. The heuristic based algorithm schedules RBs to the URLLC users before

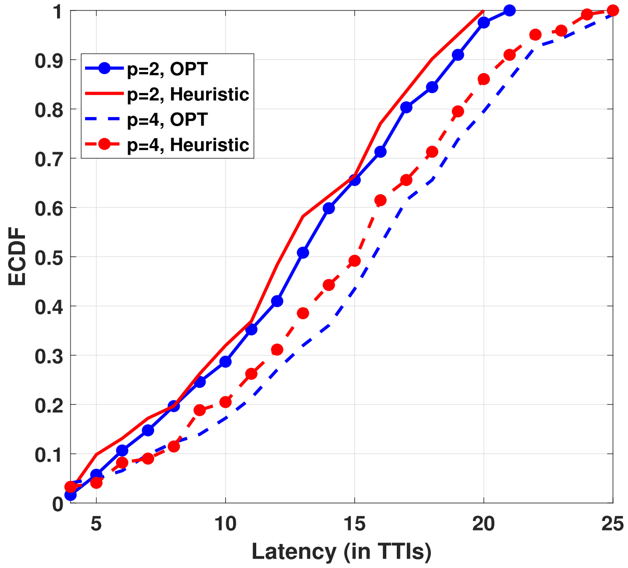


FIGURE 5. Cumulative distribution of URLLC latency for different TTI assignment strategies at the URLLC packet arrival rate of λ = 2.12 packets/frame and $L = 10$ and $K = 5$.

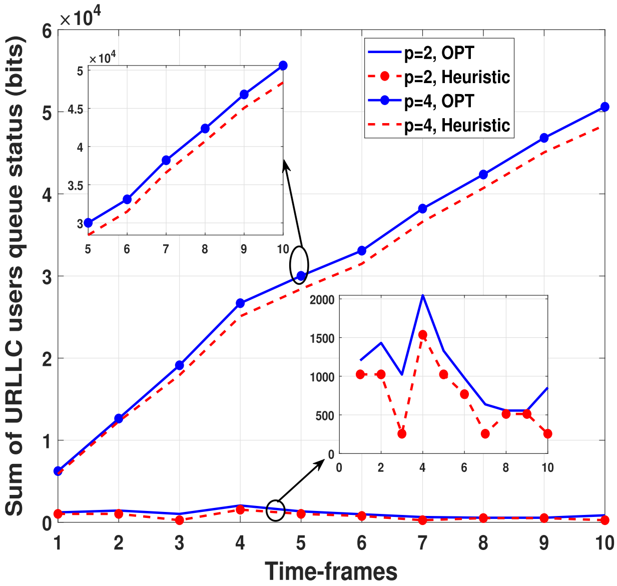


FIGURE 6. Sum of URLLC users queues on every frame for different TTI assignment strategies (p) at the URLLC packet arrival rate of λ = 6.36 packets/frame and $L = 10$ and $K = 10$.

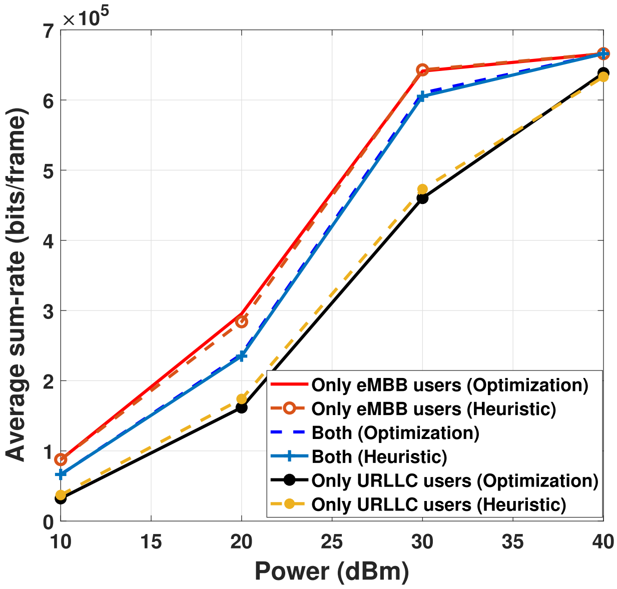


FIGURE 7. Average sum rate of eMBB, URLLC and total users with varying BS powers at 2 scheduling TTIs gap (i.e., $p = 2$) and $L = 10$ and $K = 5$.

eMBB users by giving them priority, but, the optimization method schedules RBs to users in any TTI of the considered interval range by following the constraints. This is the reason for heuristic algorithm to achieve the better performance in terms of latency compared to the optimization method.

The results in Fig. 4 and Fig. 5 show the trade off between the minimum data rate of eMBB users and the latency of URLLC users. Using the optimization based scheduling process, the minimum data rate requirement for eMBB users is achieved, while providing a good latency-related performance for URLLC users. In contrast, the heuristic scheduling scheme improves the performance of URLLC users in terms of latency, but it fails to achieve the minimum data rate for eMBB users.

Fig. 6 shows the sum of the queue status (i.e., unscheduled packets $(\zeta_u^{(l)})$) of URLLC users on every frame after executing the scheduling process using the proposed algorithms for different p values. It can be observed that both optimal and heuristic are vacating the URLLC packets for low values of "p", while this is not the case when "p" increases. Further, as can be seen from the results, the sum of unscheduled packets is increased with $p = 4$ compared to $p = 2$ as expected. By considering the assignment of RBs to URLLC users in the large TTI gaps (i.e., high values of p), obviously reduce the number of resources for URLLC users and leads to huge number of unscheduled packets in the queues of URLLC users.

In Fig. 7, we show the average sum-rate of all users obtained by using the proposed optimization-based scheduling scheme and the heuristic scheme. Specifically, we show the results by considering the three different scenarios of users existence in the cell: (i) all the distributed users belonging to eMBB service, (ii) all the distributed users belonging to URLLC service, and (iii) the presence of both the eMBB and URLLC users. From results, it is noticed that the performance of the heuristic scheme is almost as same as the performance of the optimization-based scheduling scheme (i.e., the performance gap of the two algorithms is negligible). The resulting average sum-rate using the heuristic method is a tight upper bound when all the active users in the network belong to the URLLC service. On the contrary, the resulting average sum rate using the optimization method is a tight upper bound when all the active users in the network are from the eMBB service. Also, we observe that the average sum-rate of all users increases with the BS power as expected. As the BS power increases, the received SNR at the user

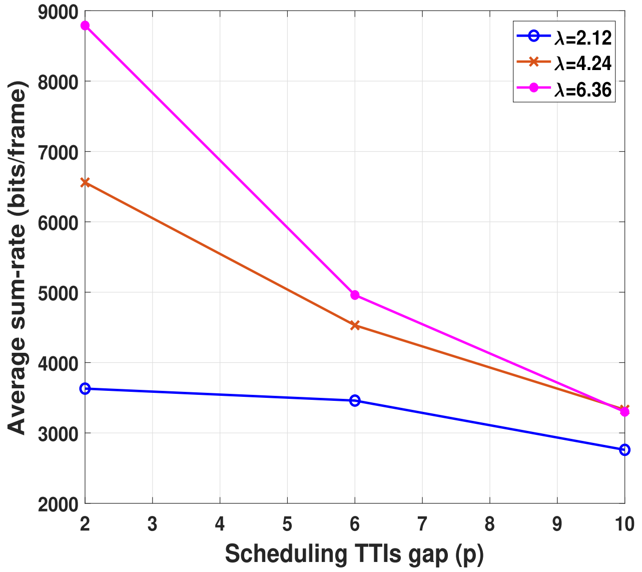


FIGURE 8. Average sum rate of URLLC with different URLLC users scheduling TTI gaps and packet arrival rates at 40 dBm of BS power (using the scheduling algorithm with constraint (C3)).

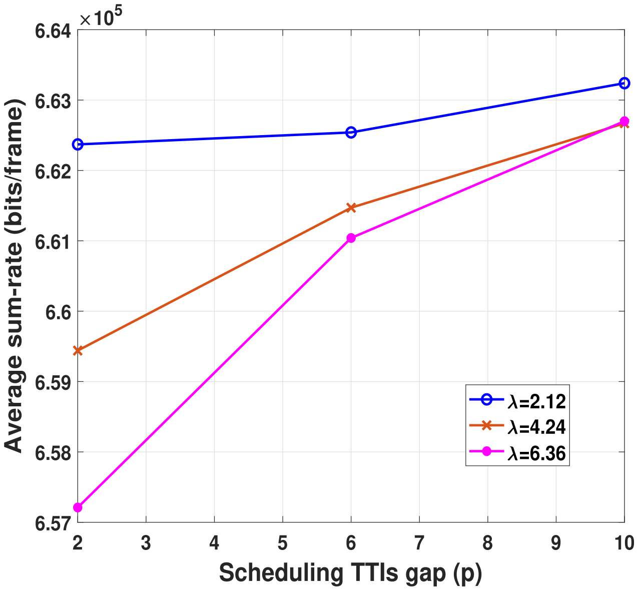


FIGURE 9. Average sum rate of eMBB with different URLLC users scheduling TTI gaps and packet arrival rates at 40 dBm of BS power (using the scheduling algorithm with constraint (C3)).

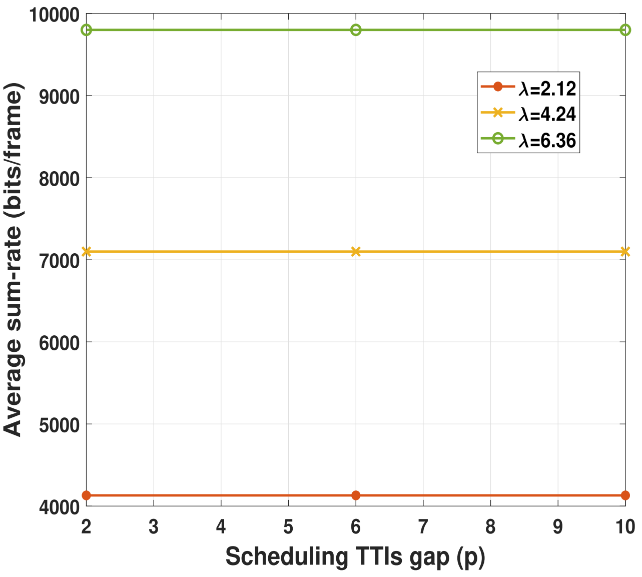


FIGURE 10. Average sum rate of URLLC with different URLLC users scheduling TTI gaps and packet arrival rates at 40 dBm of BS power (using the scheduling algorithm with a new constraint (C3a)).

increases so that the respective user chooses the higher MCS which subsequently increases the sum-rate of the user. Further, the following outcomes are observed for the three different existence scenarios: (i) when all the active users in the cell belong to eMBB service, the resulting average sum-rate performance of users is the tight upper bound, (ii) if all the active users in cell are associated with the URLLC service, then the resulting average sum-rate performance of users is the tight lower bound, and (iii) The resulting average sum-rate performance of users is lower than the performance of all eMBB users and higher than the performance of all URLLC users when the active users in cell belong to both types of services. The reliability constraint is more strict for URLLC service, therefore, in order to ensure the transmission with high success probability, the URLLC users select the lower MCS compared to the eMBB users. This is the main reason for the aforementioned results, shown in Fig. 7.

In Fig. 8 and Fig. 9, we illustrate the average sum-rate of the eMBB users and URLLC users, respectively, achieved by the proposed scheduling algorithm for different scheduling TTI gaps (i.e., different p values) and packets arrival rate of URLLC users. As can be seen from results, it is clear that the average sum-rate of URLLC users decreases by increasing the value of p. In contrast, the average sum-rate of the eMBB users increases by increasing the value of p. Specifically, this effect is very dominant at the higher packets arrival rate. This is happened due to the constraint presented in the problem that allocates only an RB to the URLLC users for every p TTIs. Obviously, this condition favors improving the sum-rate of eMBB users by assigning a higher number of RBs. Due to the allocation of less number of RBs instead of the required number of RBs, the unscheduled data packets are stacked in the queues for the longer period. For example, assume that a URLLC user is generating the data packets with the arrival rate of 8 (packets/frame). According to the arrival rate, a total 8 number of RBs are required to vacate the complete packets, which means that for every 10 TTIs, at least 4 RBs should be assigned to the user in the case of $5ms$ transmission latency. But, the constraint $(C3)$ allocates only 1 RB for every 10 TTIs (i.e., in total 2 RBs), which leads to 6 unscheduled packets in the queue, these packets transmit in the next frame. This process continues till the end of the frames that cause to the stacking of the high number of unscheduled packets in the queues.

In order to avoid the aforementioned issues and to vacate the complete available queues for URLLC users within the

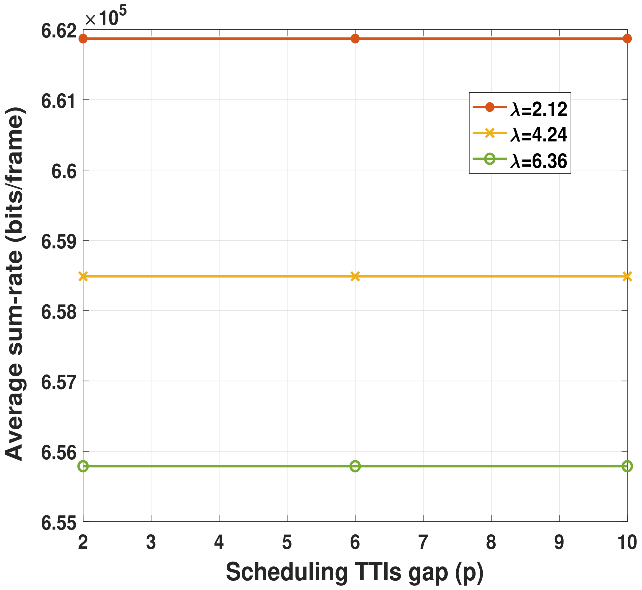


FIGURE 11. Average sum rate of eMBB with different URLLC users scheduling TTI gaps and packet arrival rates at 40 dBm of BS power (using the scheduling algorithm with a new constraint (C3a)).

latency time, the constraint (C3) can be modified as follows,

$$\sum_{f=1}^{F}\sum_{t=kp+1}^{kp+p} x_{t,f}^{u} \geq \beta; k = 0, 1, 2, 3.., k_{max}; \quad \forall u \in \mathcal{U}_2 \quad \text{(C3a)}$$

where $\beta = \max(\lceil \frac{pk_{max}}{N} \rceil, 1)$ represents the needed number of RBs for every p TTIs. Now, the results with the new constraint are illustrated in Fig. 10 and Fig. 11. The constant average sum-rates are observed for eMBB and URLLC users for all the URLLC latency requirements in the results. Further, the URLLC users achieve higher average sum-rates with the increase in the URLLC arrival rates, while the average sum-rate of eMBB users decreases with the increase in the URLLC arrival rates. The results confirm that the scheduling algorithm with the new constraint specifically useful for the URLLC users to vacate the queues within the provided scheduling TTIs.

V. CONCLUSION

In this paper, we have proposed the slice-aware RAN radio resource allocation mechanism for the dynamic multiplexing of eMBB and URLLC users on the same radio resources. The resource allocation problem was formulated as an AMC based resource optimization problem to maximize the sum-rate of the total network while satisfying the heterogeneous requirements of the users from two services. The formulated problem is a combinatorial mixed-integer non-linear programming optimization problem, which is very hard to solve in polynomial time. By relaxing the intractability of AMC and the binary constraint, the optimization problem was transformed into a continuous linear program, which was subsequently solved by using the standard CVX tool. In addition, we proposed a low-complexity heuristic scheme that significantly can reduce the computation time. Through simulation results, we show the trade off between the minimum rate requirement of eMBB users and latency requirement of URLLC users. The optimization-based scheduling algorithm outperforms the heuristic-based scheduling algorithm in terms of providing the minimum-rate to all eMBB users. In contrast, the heuristic algorithm outperforms the optimization algorithm in terms of latency and vacating the queues of URLLC users. Furthermore, the simulation results show that the overall sum-rate performances of the optimization-based scheduling and heuristic schemes are almost the same.

The proposed framework can be easily adaptable for the different time-frequency grid granularity; hence, latest numerologies proposed in [2] for the time-frequency split should be fully compatible with the proposed technique.

ACKNOWLEDGMENT

This article was presented in part at the 2019 IEEE International Workshop on Computer Aided Modeling and Design of Communication Links and Networks.

REFERENCES

[1] P. K. Korrai, E. Lagunas, S. K. Sharma, S. Chatzinotas, and B. Ottersten, "Slicing based resource allocation for multiplexing of eMBB and URLLC services in 5G wireless networks," in *Proc. IEEE 24th Int. Workshop Comput. Aided Model. Des. Commun. Links Netw. (CAMAD)*, Limassol, Cyprus, Sep. 2019, pp. 1–5.

[2] S. E. Elayoubi, S. B. Jemaa, Z. Altman, and A. Galindo-Serrano, "5G RAN slicing for verticals: Enablers and challenges," *IEEE Commun. Mag.*, vol. 57, no. 1, pp. 28–34, Jan. 2019.

[3] *Study on New Radio (NR) Access Technology Physical Layer Aspects*, document TR38.802v14.0.0, 3GPP, Mar. 2017.

[4] *ITU-R M.[IMT-2020.TECH PERF REQ]-Minimum Requirements Related to Technical Performance for IMT-2020 Radio Interface(s)*, document ITU-R M.2410-0, International Telecommunication Union-Recommendations, Nov. 2017.

[5] NGMN Alliance. *Description of Network Slicing Concept*. Accessed: Apr. 5, 2019. [Online]. Available: https://www.ngmn.org/fileadmin/user upload/160113 NetworkSlicingv10.pdf

[6] P. Popovski, K. F. Trillingsgaard, O. Simeone, and G. Durisi, "5G wireless network slicing for eMBB, URLLC, and mMTC: A communication-theoretic view," *IEEE Access*, vol. 6, pp. 55765–55779, 2018.

[7] *Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond*, document ITU-R M.2083-0, International Telecommunication Union, Feb. 2015.

[8] J. van de Belt, H. Ahmadi, and L. E. Doyle, "Defining and surveying wireless link virtualization and wireless network virtualization," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, pp. 1603–1627, 3rd Quart., 2017.

[9] S. Vassilaras, L. Gkatzikis, N. Liakopoulos, I. N. Stiakogiannakis, M. Qi, L. Shi, L. Liu, M. Debbah, and G. S. Paschos, "The algorithmic aspects of network slicing," *IEEE Commun. Mag.*, vol. 55, no. 8, pp. 112–119, Aug. 2017.

[10] *New WID on New Radio Access Technology*, document RP-170855, RAN75, 3GPP, Dubrovnik, Croatia, Mar. 2017. Accessed: Jun. 18, 2018. [Online]. Available: http://www.3gpp.org/ftp/TSGRAN/TSGRAN/TSGR75/Docs/RP-170855.zip

[11] M. I. Kamel, L. B. Le, and A. Girard, "LTE wireless network virtualization: Dynamic slicing via flexible scheduling," in *Proc. IEEE 80th Veh. Technol. Conf. (VTC-Fall)*, Vancouver, BC, Canada, Sep. 2014, pp. 1–5.

[12] M. Hu, Y. Chang, Y. Sun, and H. Li, "Dynamic slicing and scheduling for wireless network virtualization in downlink LTE system," in *Proc. 19th Int. Symp. Wireless Pers. Multimedia Commun. (WPMC)*, Shenzhen, China, Nov. 2016, pp. 153–158.

[13] S. Parsaeefard, V. Jumba, M. Derakhshani, and T. Le-Ngoc, "Joint resource provisioning and admission control in wireless virtualized networks," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, New Orleans, LA, USA, Mar. 2015, pp. 2020–2025.

[14] Y. Zhang, L. Zhao, D. Lopez-Perez, and K.-C. Chen, "Energy-efficient virtual resource allocation in OFDMA systems," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Washington, DC, USA, Dec. 2016, pp. 1–6.

[15] C.-P. Li, J. Jiang, W. Chen, T. Ji, and J. Smee, "5G ultra-reliable and low-latency systems design," in *Proc. Eur. Conf. Netw. Commun. (EuCNC)*, Oulu, Finland, Jun. 2017, pp. 1–5.

[16] G. Pocovi, K. I. Pedersen, and P. Mogensen, "Joint link adaptation and scheduling for 5G ultra-reliable low-latency communications," *IEEE Access*, vol. 6, pp. 28912–28922, 2018.

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

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

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

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

[21] Q. Liu and C. W. Chen, "Smart downlink scheduling for multimedia streaming over LTE networks with hard handoff," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 11, pp. 1815–1829, Nov. 2015.

[22] X. Lin, J. Li, R. Baldemair, T. Cheng, S. Parkvall, D. Larsson, H. Koorapaty, M. Frenne, S. Falahati, A. Grövlen, and K. Werner, "5G new radio: Unveiling the essentials of the next generation wireless access technology," 2018, *arXiv:1806.06898*. [Online]. Available: http://arxiv.org/abs/1806.06898

[23] T. Innovations, "LTE in a nutshell," White Paper, 2010.

[24] Y. Zaki, *Future Mobile Communications: LTE Optimization and Mobile Network Virtualization*, vol. 1. Berlin, Germany: Springer, 2012.

[25] D. Lopez-Perez, A. Ladanyi, A. Juttner, H. Rivano, and J. Zhang, "Optimization method for the joint allocation of modulation schemes, coding rates, resource blocks and power in self-organizing LTE networks," in *Proc. Proc. IEEE INFOCOM*, Shanghai, China, Apr. 2011, pp. 111–115.

[26] E. Hossain, M. Rasti, and L. B. Le, *Radio Resource Management in Wireless Networks: An Engineering Approach*. Cambridge, U.K.: Cambridge Univ. Press, 2017.

[27] A. L. Yuille and A. Rangarajan, "The concave-convex procedure (CCCP)," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 2, 2002, pp. 1033–1040.

[28] G. R. Lanckriet and B. K. Sriperumbudur, "On the convergence of the concave-convex procedure," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 1759–1767.

[29] M. Grant and S. Boyd. (2011). *CVX: MATLAB Software for Disciplined Convex Programming*. [Online]. Available:http://cvxr.com/cvx

[30] B. Khodapanah, A. Awada, I. Viering, J. Francis, M. Simsek, and G. P. Fettweis, "Radio resource management in context of network slicing: What is missing in existing mechanisms?" in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Marrakesh, Morocco, Apr. 2019, pp. 1–7.

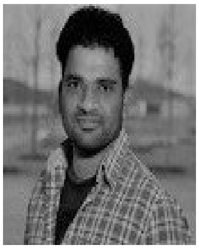
[31] A. K. Bairagi, M. S. Munir, M. Alsenwi, N. H. Tran, and C. S. Hong, "A matching based coexistence mechanism between eMBB and uRLLC in 5G wireless networks," in *Proc. 34th ACM/SIGAPP Symp. Appl. Comput. (SAC)*, 2019, pp. 2377–2384.

[32] P. Gahinet, A. Nemirovski, A. J. Laub, and M. Chilali, *LMI Control Toolbox Users Guide*. Natick, MA, USA: MathWorks, 1995.

[33] M. Mohseni, S. A. Banani, A. W. Eckford, and R. S. Adve, "Scheduling for VoLTE: Resource allocation optimization and low-complexity algorithms," *IEEE Trans. Wireless Commun.*, vol. 18, no. 3, pp. 1534–1547, Mar. 2019.

[34] J. Zhang and J. G. Andrews, "Adaptive spatial intercell interference cancellation in multicell wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 28, no. 9, pp. 1455–1468, Dec. 2010.

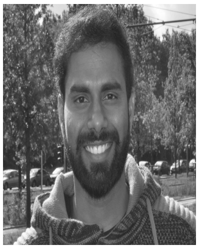
[35] N. Seifi, M. Matthaiou, and M. Viberg, "Coordinated user scheduling in the multi-cell MIMO downlink," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Prague, Czech Republic, May 2011, pp. 2840–2843.



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

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

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



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

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

BJÖRN OTTERSTEN (Fellow, IEEE) was born in Stockholm, Sweden, in 1961. He received the M.S. degree in electrical engineering and applied physics from Linkoping University, Linkoping, Sweden, in 1986, and the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, USA, in 1990. He has held research positions with the Department of Electrical Engineering, Linkoping University; the Information Systems Laboratory, Stanford University; the Katholieke Universiteit Leuven, Leuven, Belgium; and the University of Luxembourg, Luxembourg. From 1996 to 1997, he was the Director of research with ArrayComm, Inc., a start-up in San Jose, CA, based on his patented technology. In 1991, he was appointed as a Professor of signal processing with the Royal Institute of Technology (KTH), Stockholm, Sweden. From 1992 to 2004, he was the Head of the Department for Signals, Sensors, and Systems, KTH. From 2004 to 2008, he was the Dean of the School of Electrical Engineering, KTH. He is currently the Director for the Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg.

He is a Fellow of EURASIP. He was a recipient of the IEEE Signal Processing Society Technical Achievement Award, in 2011, and the European Research Council advanced research grant twice, from 2009 to 2013 and from 2017 to 2022. He has coauthored journal articles that received the IEEE Signal Processing Society Best Paper Award, in 1993, 2001, 2006, and 2013, and seven IEEE conference papers best paper awards. He is currently a member of the editorial boards of *EURASIP Signal Processing Journal*, *EURASIP Journal of Advances Signal Processing*, and *Foundations and Trends of Signal Processing*. He has served as an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING and the Editorial Board of the *IEEE Signal Processing Magazine*.