Slicing based Resource Allocation for Multiplexing of eMBB and URLLC Services in 5G Wireless Networks

PraveenKumar Korrai, Eva Lagunas, Shree Krishna Sharma, Symeon Chatzinotas and Björn Ottersten
Interdisciplinary Centre for Security, Reliability and Trust
University of Luxembourg, Luxembourg
E-mail:{praveen.korrai, eva.lagunas, shree.sharma, symeon.chatzinotas, and bjorn.ottersten} @uni.lu

***Abstract*—The next generation of wireless networks is intended to accommodate two major services: enhanced mobile broadband (eMBB), and ultra-reliable and low-latency communications (URLLC). The eMBB applications require higher data rates while URLLC applications require a stringent latency and high transmission success probability (i.e., reliability). The multiplexing of eMBB and URLLC services on the same network infrastructure leads to a challenging resource optimization problem. In this paper, we formulate the network slicing problem in the context of time-frequency resource blocks (RBs) allocation to the wireless system consisting of eMBB and URLLC users. In particular, we address the sum-rate maximization problem subject to latency and slice isolation constraints while assuring certain reliability requirements with the use of adaptive modulation coding (AMC). We relax the mathematical intractability of AMC and binary assignment variable and show the effectiveness of the proposed approach through numerical simulations.**

***Index Terms*—Radio resource allocation, eMBB, URLLC, Resource slicing.**

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 [1]. To further improve our cities, living environment and industries, the fifth generation (5G) of wireless networks is required to provide support for a large variety of services and applications. Towards this achievement, besides the enhanced mobile broadband (eMBB) service, supporting today's high broadband traffic, the upcoming 5G New Radio (NR) has to support a new service: ultra-reliable and low latency communications (URLLC), described by the stringent requirements in terms of high reliability and low latency [2], [3]. However, accommodating these different wireless services in the same physical network while assuring their potent coexistence is a major challenge.

Recently, network slicing has emerged as a promising technology for allocating resources to different wireless services with diverse quality-of-service (QoS) needs [4]. In particular, the network slice is created on top of the same physical network such that it is not distinct from the separate physical network [5]. Thus, the created each slice acts as a dedicated network for a specific service. Moreover, this slicing is executed both on the Random Access Network (RAN) and the Core Network (CN). While the challenges of creating slices in CN have been extensively addressed [6], the RAN slicing is still posing challenges due to the radio-resource-demanding and multi-services environment.

Due to its flexibility in radio resources allocation, Orthogonal Frequency Division Multiple Access (OFDMA) is envisaged in the current 5G proposals for the downlink (DL) wireless systems [7]. An intuitive approach of creating a RAN slice is the dynamic assignment of resource blocks (RBs), which is defined in 3GPP 5G-NR [8] as minimum time-frequency resource that can be assigned to a particular user.

Recently, RAN radio resource allocation has attracted significant research attention. A resource scheduling technique to slice an LTE network into multiple virtual networks (VNs) to provide services for different service providers has been investigated in [9]. The work in [10] proposed a new slicing and scheduling mechanism for wireless virtual networks, which dynamically allocates a specific number of RBs to each VN to provide services to its users. In [11], the authors have studied the admission control and resource provisioning problems for OFDMA based wireless VNs. The authors of [12] have proposed the energy-efficient sub-carrier and power allocation strategy with wireless network virtualization (WNV) for a single-cell OFDMA system. Although the aforementioned works have assumed some constraints on resources assignment to maintain isolation between the slices, they did not consider the end-to-end QoS requirements such as latency and reliability.

As aforementioned, the 5G NR has to support the multiple number of services and a huge variety of applications. According to [13], the two main services to be supported in 5G wireless networks are eMBB and URLLC. In this context, the multiplexing of eMBB and URLLC traffics on the same wireless network infrastructure has recently received a significant research attention in both academia and industry. Through a simplified queuing analysis, the authors of [14] have shown that the efficient multiplexing of eMBB and URLLC traffic (in both the frequency and time domains) increases the overall resource efficiency of a wireless system. In [15], a punctured scheduling mechanism has been introduced for the transmission of latency critical traffic on a shared channel with the eMBB traffic.

However, the works [14, 15] have not provided the so-

lution for how to efficiently allocate the wireless resources between URLLC and eMBB services while assuring their stringent QoS requirements.To address this problem, QoS aware scheduling functionalities need to be invoked. Further, reliable data transmissions can be achieved to some extent by choosing a suitable modulation and coding scheme (MCS) such that lower block error rate (BLER) is obtained. To this end, different from the existing works, we propose a radio resource allocation mechanism with an adaptive modulation coding (AMC) scheme for the dynamic multiplexing of eMBB and URLLC services in 5G wireless networks. This resource allocation problem is formulated as an optimization problem to maximize the sum rate of all users, while satisfying the latency requirement of URLLC users. In this AMC based optimization problem, each RB's achievable data rate is estimated by the selected MCS instead of Shannon rate formula. Further, we utilize different SNR levels for eMBB and URLLC users to select the MCS in accordance to their target BLER.

The remainder of the paper is organized as follows. The system model is described in Section II. Problem formulation is presented in Section III. Section IV provides the numerical evaluations and discussions. Finally, the conclusions are drawn in Section V.

II. System Model

We consider the DL OFDMA scenario of a single-cell wireless cellular network, where a base station (BS) is located at the center of the cell, and user equipment (UEs) are distributed uniformly across network as shown in Fig. 1. We further assume that the distributed UEs are associated with different types of services such as eMBB and URLLC (i.e., means that different services are existed in the network). The BS serves all the UEs in the cell, indexed by $\mathcal{U} = \{1, 2, ...M\}$, 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 $f = \{1, 2, ...F\}$ and the time dimension is divided into time-slots indexed by $t = \{1, 2, ...N\}$ with the length of 0.5 ms as shown in Fig. 2. An RB is the minimum assignment resource, which consists of 7 OFDM symbols in a time-slot of 0.5 ms and 12 consecutive sub-carriers (for a complete bandwidth of 180 KHz) [16].

We assume that the channel state information (CSI) is perfectly available at the BS from all the users belonging to the different services, and also assume that the power P_{max} is equally allocated over all sub-bands (i.e., allocated power to each sub-band is $P = P_{max}/F$).Then, the signal-to-noise ratio (SNR) of user u on sub-band f at the time slot t is computed as

$$\gamma^u_{t,f} = \frac{P|h^u_{t,f}|^2 d^{-\alpha}_{BS,u}}{\sigma^2}, \tag{1}$$

where $h^u_{t,f}$ represents the channel gain of user u on a sub-band f at the time slot 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.

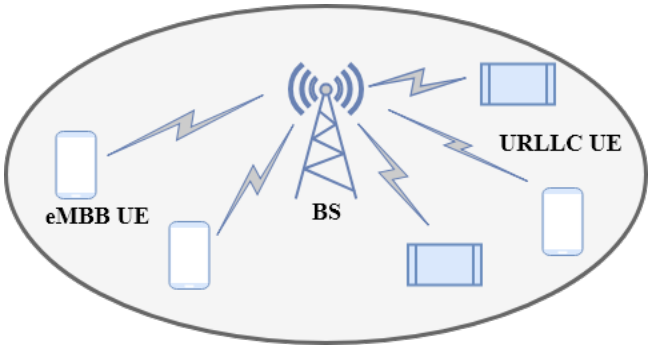

Fig. 1. Single-Cell wireless network with eMBB and URLLC users

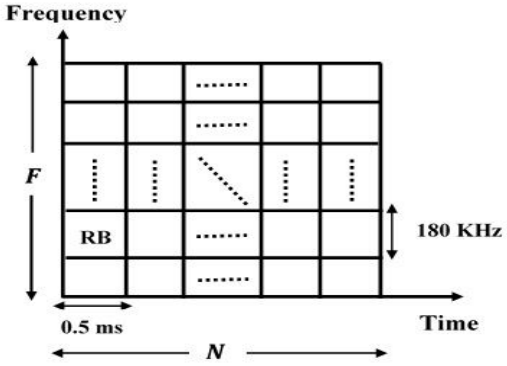

Fig. 2. Time and frequency (2D) radio resources

III. Problem Formulation and Proposed Solution

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. AMC method allows the wireless system to choose the appropriate MCS according to the received channel quality. Based on the target block error rate (BLER) or probability of error, and the received channel quality indicator (CQI) feedback from the user, the minimum SNR threshold is set to obtain the appropriate MCS. In our study, M different MCSs are considered and the corresponding SNR levels of MCS for distinct BLER targets are provided in Table I [17].

Using the MCS, the bit rate of user u operating in sub-band f at time slot t can be expressed as

$$R^u_{t,f} = B \cdot T \cdot \mathcal{F}(\gamma) \tag{2}$$

where B 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 I according to the received SNR. During the scheduling round, 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 when the RB is not assigned to any user. Then, the binary constraint is mathematically written as

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

TABLE I
MODULATION AND CODING SCHEMES

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

The bit rate of each user u, D_u is computed as

$$D_u = \sum_{t=1}^{N} \sum_{f=1}^{F} x_{t,f}^u R_{t,f}^u \tag{3}$$

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

$$R = \sum_{u \in \mathcal{U}_1} D_u + \sum_{u \in \mathcal{U}_2} D_u \tag{4}$$

We maximize the above objective function subject to the constraints as follows. The constraint (C2) assures that an RB can only be allocated to a single user belongs to one service in a time slot (i.e., orthogonality constraint). The next constraint (C3) guarantees the latency requirement of URLLC users. When the URLLC user is scheduled (i.e., $u \in \mathcal{U}_2$), it must receive at least one RB for the every N time-slots to maintain the latency requirement of the user. The main objective of the proposed optimization problem is to maximize the overall sum-rate of users associated with eMBB and URLLC services through performing a dynamic resource blocks (RBs) allocation subject to a set of constraints. Mathematically, the optimization problem is expressed as

$$\textbf{P1:} \quad \max_{\{x_{t,f}^u\}} \quad R \tag{5}$$

subject to

$$x_{t,f}^u \in \{0, 1\} \tag{C1}$$

$$\sum_{u \in \mathcal{U}_1} \sum_{u \in \mathcal{U}_2} x_{t,f}^u = 1; \forall\ t, f \tag{C2}$$

$$\sum_{f=1}^{F} \sum_{t=kN+1}^{kN+N} x_{t,f}^u \geq 1; k = 0, 1, 2, 3..; u \in \mathcal{U}_2 \tag{C3}$$

However, the presented function $\mathcal{F}(.)$ in (5) is a step-wise function that makes the optimization problem mathematically intractable and complex to solve. In order 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

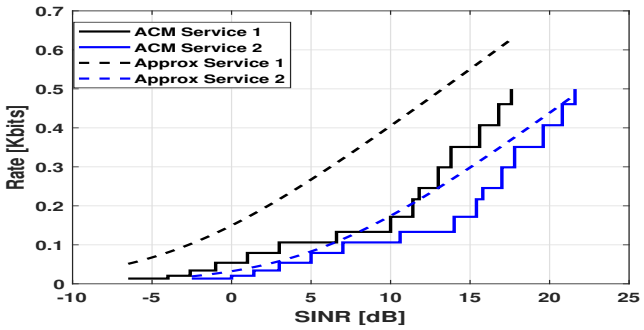

Fig. 3. Rate of eMBB, URLLC services using ACM scheme and approximated rate functions

$$\mathcal{F}_E(\gamma_{t,f}^u, \beta_E) = \log_2(1 + \frac{\gamma_{t,f}^u}{\Gamma_E}), \tag{6}$$

$$\mathcal{F}_U(\gamma_{t,f}^u, \beta_U) = \log_2(1 + \frac{\gamma_{t,f}^u}{\Gamma_U}) \tag{7}$$

where $\Gamma_E = \frac{-\ln(5\beta_E)}{1.5}$ and $\Gamma_U = \frac{-\ln(5\beta_U)}{1.5}$ [18] represent the SNR gaps, β_E and β_U represent the target BLER for eMBB and URLLC users, respectively. The proposed approximate functions are compared to the values achieved with AMC Table I in Fig. 3.

Now, the bit rate of each RB can be written as

$$r_{t,f}^u = B \cdot T \cdot\ \mathcal{F}_a(\gamma, \beta_a),\ a \in \{E, U\} \tag{8}$$

Using (8), the objective function sum-rate of all users can be reformulated as

$$V = \sum_{u \in \mathcal{U}_1} \sum_{t=1}^{N} \sum_{f=1}^{F} x_{t,f}^u r_{t,f}^u + \sum_{u \in \mathcal{U}_2} \sum_{t=1}^{N} \sum_{f=1}^{F} x_{t,f}^u r_{t,f}^u \tag{9}$$

The optimization problem (P1) is now reformulated as

$$\textbf{P2:} \quad \max_{\{x^u_{t,f}\}} \quad V \tag{10}$$

$$\text{subject to}$$

$$(C1), (C2), \text{and } (C3) \tag{11}$$

The problem P2 is combinatorial due to the constraint C1 and non-convex due to the objective function and the constraint (C1). To avoid the combinatorial nature of P1, the assignment variable $x^u_{t,f}$ is relaxed to $0 \leq x^u_{t,f} \leq 1$. After relaxing the binary constraint, the problem becomes a continuous linear program that can be solved by the standard tools. Once P2 is solved, we apply hard thresholding to recover the binary variable $x^u_{t,f}$. Further, the proposed solution optimizes the resource allocation across both frequency and time simultaneously, covering the complete frame. The final output shows in each scheduling round which users should be served on which RB. The proposed solution provides $< 10^{-3}$ optimality gap with respect to the linear programming relaxation problem.

IV. Numerical Evaluations

In this section, the proposed slicing based scheduling mechanism is evaluated in the downlink single-cell OFDMA scenario. To analyze the proposed mechanism, we consider 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.

A. Simulation Environment

In this work, we concentrate on the resource allocation for a downlink wireless system. A BS is deployed at the center of the cell coverage area with the radius of 250m. L eMBB users and K URLLC users are distributed randomly within the cell coverage area. Further, we assume that the channel between the BS and the user follows a Nakagami-m fading model. Also, the path loss exponent (α) is set to 4. Our simulations are executed for a frame of 10 ms (i.e., 20 time-slots) with a bandwidth of 20 MHz consists of 100 RBs. Each RB comprises of 12 sub-carrier each with a spacing of 15 KHz. The bandwidth of each RB is 180 KHz. Further, we assume white Gaussian noise power on each sub-band is 10^{-11}W. Also, consider that the target BLER of URLLC users (i.e., $\beta_U = 10^{-3}$) is lower than the eMBB users (i.e., $\beta_E = 10^{-1}$). The complete set of simulation parameters is provided in Table II. All the simulations were performed and averaged over 100 independent Monte-Carlo runs.

TABLE II
Simulation Parameters

Parameter	**Value**
BS Transmit power (P_{max})	10, 20, 30, 40 dBm
path loss exponent (α)	4
Channel	Nakagami-m fading model
Number of time-slots (N)	20
Each time-slot length (T)	0.5 ms
Total length of time-frame	10ms
Number of OFDM symbols per time-slot	7
RBs per time-slot (F)	100
Number of sub-carriers per RB	12
Each sub-carrier length	15 KHz
Each RB's bandwidth	180 KHz
Carrier Bandwidth	20 MHz
Cell radius	250 m
Number of users (M)	5, 10
Number of eMBB users (L)	3, 5
Number of URLLC users (K)	2, 5

B. Results and Discussions

Fig. 4 shows the sum-rate of all users achieved by the proposed resource allocation mechanism for different values of BS transmit power. From the results, it is observed that the sum-rate increases with the BS power as expected. This result occurs due to the fact that as the BS increases power the received SNR improves at the user so that the respective user may select the higher order MCS that increases the sum-rate of user. Also we observe the following results for three different scenarios: (i) the resulting sum-rate of users provides the upper bound when all the active users in the cell are associated with the eMBB services, (ii) In contrast to the first scenario, the resulting sum-rate of users provides the lower bound when all the active users are associated with the URLLC service, and (iii) The resulting total sum-rate of the users is lower than the total sum-rate of all eMBB users and higher than the total sum-rate of all URLLC users when the active users in cell are associated with the two types of services. The reliability constraint is more strict for URLLC users, so that in order to maintain the transmission with high success probability, the URLLC users choose the lower MCS compared to the eMBB users. This is the major reason for the aforementioned results.

In Fig. 5, we plot the sum-rate of users with the latency requirement of the URLLC users. It is noticed from the results that the sum-rate is improved minimally by increasing the latency time of URLLC service. In particular, we observe from sub-plots that the sum-rate of URLLC users decreases as increases the latency time of URLLC users, and contrastingly the sum-rate of eMBB users increases as raises the latency time of URLLC users. In results, the sum-rate is maximized at the high latency time when both eMBB and URLLC users are active in the cell. These results are occurred due to the allocation of less number of RBs to URLLC users and more number of RBs to eMBB users at the high latency time of URLLC users. Further, the eMBB users utilize the higher MCS compared to URLLC users, that in turn improves the total sum-rate of users.

V. Conclusions

In this paper, we have proposed a slicing based resource allocation technique for the efficient multiplexing of eMBB and URLLC users on the same radio resources. Furthermore, an AMC based resource optimization was utilized while selecting the appropriate MCS for improving the transmission reliability. This radio resource allocation problem was formulated as a combinatorial integer non-linear programming optimization

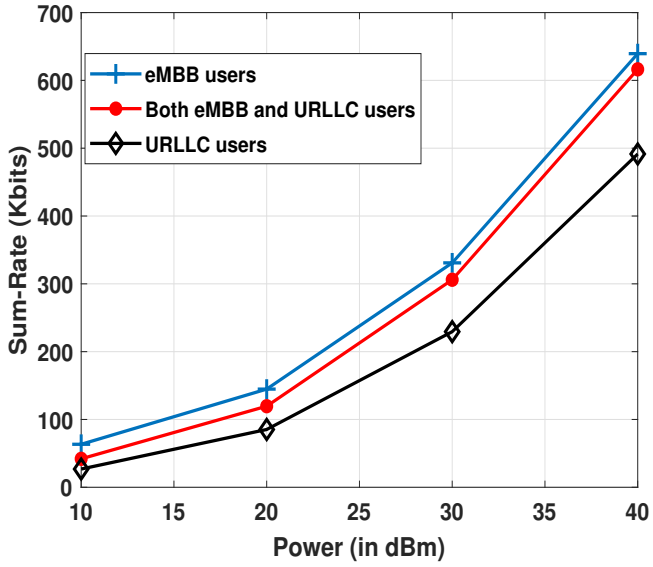

Fig. 4. Sum rate of eMBB, URLLC and total users with varying BS powers and 1 ms URLLC latency requirement. A total 5 number of users (i.e., 3 eMBB and 2 URLLC users) are considered for this simulation.

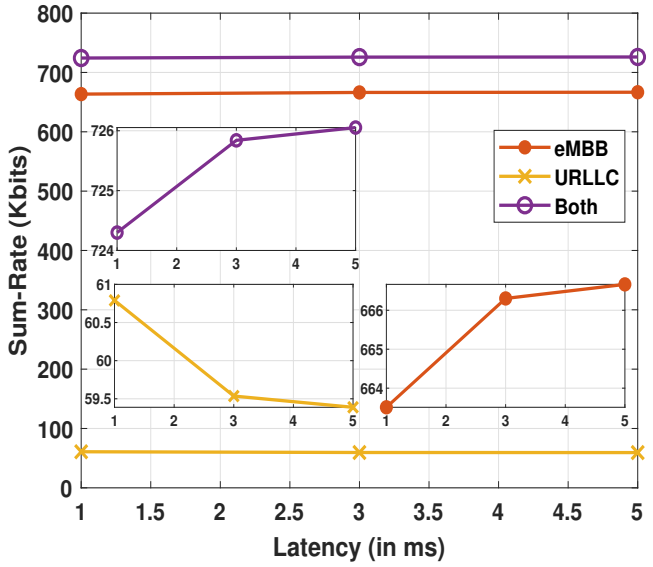

Fig. 5. Sum rate of URLLC and total users with different URLLC latency requirements at 40 dBm of BS power. A total 10 number of users (i.e., 5 eMBB and 5 URLLC users) are considered for this simulation

problem, which is very hard to solve in polynomial time. Later, we relaxed the AMC and binary constraints, and transformed the problem into continuous linear program, which was solved by using the standard CVX tool. The simulation results show that by allocating the resources dynamically for eMBB and URLCC users, the overall sum-rate of users is higher than the only URLLC users sum-rate and a little lower than the only eMBB users sum-rate. Also, a minimal total sum-rate improvement is observed at the high latency time of URLLC users.

ACKNOWLEDGMENT

This work has received funding from Luxembourg National Research Fund (FNR) under the AFR grant for the PhD project LACLOCCN (AFR grant reference No 12561031).

REFERENCES

[1] S. E. Elayoubi, S. B. Jemaa, Z. Altman and A. Galindo-Serrano, "5G RAN Slicing for Verticals: Enablers and Challenges," in IEEE Communications Magazine, vol. 57, no. 1, pp. 28-34, January 2019.
[2] Study on New Radio (NR) Access Technology Physical Layer Aspects, document 3GPP TR38.802v14.0.0, Mar. 2017.
[3] P. Popovski, K. F. Trillingsgaard, O. Simeone and G. Durisi, "5G Wireless Network Slicing for eMBB, URLLC, and mMTC: A Communication-Theoretic View," in IEEE Access, vol. 6, pp. 55765-55779, 2018.
[4] J. van de Belt, H. Ahmadi and L. E. Doyle, "Defining and Surveying Wireless Link Virtualization and Wireless Network Virtualization," in IEEE Communications Surveys and Tutorials, vol. 19, no. 3, pp. 1603-1627, thirdquarter 2017.
[5] S. Vassilaras et al., "The Algorithmic Aspects of Network Slicing," in IEEE Communications Magazine, vol. 55, no. 8, pp. 112-119, Aug. 2017.
[6] Z. A. Qazi, M. Walls, A. Panda, V. Sekar, S. Ratnasamy, and S. Shenker, A high performance packet core for next generation cellular networks, in Proc. Conf. ACM Special Interest Group Data Commun., 2017, pp. 348361
[7] Z. E. Ankarali, B. Pekoz, and H. Arslan, Flexible Radio Access Beyond 5G: A Future Projection on Waveform, Numerology, and Frame Design Principles, IEEE Access, 2017.
[8] RP-170855, New WID on new radio access technology, 3GPP RAN 75,Dubrovnik,Croatia,March 2017.Available online:http://www.3gpp.org/ftp/TSGRAN/TSGRAN/TSGR75/Docs/RP-170855.zip. Accessed on June 18, 2018
[9] M. I. Kamel, L. B. Le and A. Girard, "LTE Wireless Network Virtualization: Dynamic Slicing via Flexible Scheduling," 2014 IEEE 80th Vehicular Technology Conference (VTC2014-Fall), Vancouver, BC, 2014, pp. 1-5.
[10] M. Hu, Y. Chang, Y. Sun and H. Li, "Dynamic slicing and scheduling for wireless network virtualization in downlink LTE system," 2016 19th International Symposium on Wireless Personal Multimedia Communications (WPMC), Shenzhen, 2016, pp. 153-158.
[11] S. Parsaeefard, V. Jumba, M. Derakhshani and T. Le-Ngoc, "Joint resource provisioning and admission control in wireless virtualized networks," 2015 IEEE Wireless Communications and Networking Conference (WCNC), New Orleans, LA, 2015, pp. 2020-2025.
[12] Y. Zhang, L. Zhao, D. Lopez-Perez and K. Chen, "Energy-Efficient Virtual Resource Allocation in OFDMA Systems," 2016 IEEE Global Communications Conference (GLOBECOM), Washington, DC, 2016, pp. 1-6.
[13] Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond, document ITU-R M.2083-0, International Telecommunication Union (ITU), Feb. 2015.
[14] Chih-Ping Li, Jing Jiang, W. Chen, Tingfang Ji and J. Smee, "5G ultra-reliable and low-latency systems design," 2017 European Conference on Networks and Communications (EuCNC), Oulu, 2017, pp. 1-5.
[15] K. I. Pedersen, G. Pocovi, J. Steiner and S. R. Khosravirad, "Punctured Scheduling for Critical Low Latency Data on a Shared Channel with Mobile Broadband," 2017 IEEE 86th Vehicular Technology Conference (VTC-Fall), Toronto, ON, 2017, pp. 1-6.
[16] X. Lin, J. Li, R. Baldemair, T. Cheng, S. Parkvall, D. Larsson, H. Koorapaty, M. Frenne, S. Falahati, A. Gr ovlen, and K. Werner, 5G New Radio: Unveiling the essentials of the next generation wireless access technology, 2018, Available: [Online]. https://arxiv.org/pdf/1806.06898.
[17] D. Lpez-Prez, . Ladnyi, A. Jttner, 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," 2011 Proceedings IEEE INFOCOM, Shanghai, 2011, pp. 111-115.
[18] E. Hossain, M. Rasti, and L. B. Le, Radio Resource Management in Wireless Networks: An Engineering Approach. Cambridge, U.K.: Cambridge Univ. Press, 2017.