IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 10 September 2023, accepted 1 October 2023, date of publication 11 October 2023, date of current version 19 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3323931

RESEARCH ARTICLE

AI-Assisted Dynamic Frame Structure With Intelligent Puncturing Schemes for 5G Networks

MOHAMMAD REZA ABEDI[1], (Student Member, IEEE), MOHAMMAD REZA JAVAN[2], (Senior Member, IEEE), MOHSEN POURGHASEMIAN[1], NADER MOKARI[1], (Senior Member, IEEE), AND EDUARD A. JORSWIECK[3], (Fellow, IEEE)

[1]Department of Electrical and Computer Engineering, Tarbiat Modares University, Tehran 14115111, Iran
[2]Faculty of Electrical Engineering, Shahrood University of Technology, Shahrood 3619995161, Iran
[3]Institute for Communication Technology, TU Braunschweig, 38100 Braunschweig, Germany

Corresponding author: Mohammad Reza Javan (javan@shahroodut.ac.ir)

ABSTRACT The dynamic resource block structure (D-RBS) allows for flexible allocation of radio resources. This flexibility (potentially) enables efficient utilization of available resources and adaptability to changing network conditions. In this context, managing resource contention and optimizing allocation decisions become increasingly challenging. In this research paper, we introduce a new approach for D-RBS for re-allocation of enhanced Mobile Broadband (eMBB) and massive Machine Type Communication (mMTC) resource blocs (RBs) to URLLC users. Our scheme leverages artificial intelligence (AI) to support the three main services of 5th generation networks. To efficiently allocate resources for eMBB/mMTC and URLLC services, we propose an intelligent puncturing scheme. Additionally, we formulate an optimization problem that aims to minimize resource and transmit power usage while meeting the requirements of all three services. Since this problem is non-convex and involves multiple optimization variables, we utilize deep reinforcement learning as a solution algorithm. We then compare our proposed intelligent allocation (IA) scheme with two other schemes: random allocation (RaA) and overallocation (OA), which have lower complexity and overhead. Performance and complexity analyses are conducted in a multi-cell scenario with interference. Our results demonstrate that the IA scheme outperforms RaA and OA, achieving an energy efficiency gain of 40% and 15% respectively. However, it is worth noting that IA has a 36.3% higher complexity in terms of action selection compared to RaA and OA.

INDEX TERMS Dynamic RB structure, URLLC puncturing, eMBB and URLLC multiplexing, MA-DRL.

I. INTRODUCTION

ITU-R has categorized the network services into enhanced Mobile Broadband (eMBB), massive Machine Type Communications (mMTC), and Ultra Reliable and Low Latency communications (URLLC) services [1]. eMBB services, like video streaming, require high data rate connections (higher than 100 Mb/s); mMTC services, like Internet of Things (IoT), require large number of devices to connect to the network which would only send small data payloads; and URLLC services, like Tactile Internet (TI), require communications of small payloads with low data rates (0.1-10 Mb/s) but extremely stringent latency (∼1 ms) and very high reliability requirements [2] and [3]. Due to the heterogeneous requirements of these services, the efficient and dynamic Resource Allocation (RA) in the network is the main challenge. The traditional Resource Block Structure (RBS) definition with fixed time-frequency structure, e.g., as in Long Term Evolution (LTE) with 1 millisecond Transmission Time Interval (TTI) and 15 kHz bandwidth, is not able to support the network with these three types of services [1]. The eMMB services require more frequency resources to satisfy their data rate requirements whose latency is not of importance, and hence, the time duration of the allocated Resource Blocks (RBs) could be high, e.g., above one millisecond. On the other hand, URLLC services have

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

small packets which must be sent in short time intervals as mini-slots with the duration of 0.1 to 0.5 milliseconds, with high reliability. The mini-slot constitutes scalable numbers of Orthogonal Frequency-division Multiplexing (OFDM) symbols ranging from 14, 7, 4, and 2. In fifth generation (5G), the RBs with different TTIs and bandwidths can satisfy the data rate, latency, number of connections, and reliability requirements of the services. The notion of RB in 5G networks is defined as a time-frequency block with different time durations and bandwidths. In other words, the total time duration and bandwidth (i.e., Resource Grid (RG)), e.g., T milliseconds time duration and W Hz bandwidth, is partitioned into several RBs with different time durations, e.g., T, $\frac{T}{2}$, and $\frac{T}{4}$, and different bandwidths, e.g., W, $\frac{W}{2}$, and $\frac{W}{4}$. Note that different RBs with different time durations and bandwidths can be used for different services based on the service requirements. Although the size of the RBs is different, the location of each RB is predetermined and fixed in the RG. We call this structure as Static RBS (S-RBS) as the location of the RBs is predefined and cannot be changed. Due to the fixed location of RBs in RG, S-RBS does not provide enough flexibility to satisfy different service requirements in an efficient way. Moreover, multiplexing eMBB and mMTC with URLLC services on the same channel is another challenging problem. To overcome this challenge, 3GPP standards propose a new RA scheme, in which an arriving URLLC packet occupy the RBs that have already been allocated to eMBB or mMTC users. In other words, some RBs are punctured which may have negative impacts on eMBB or mMTC users data rate [4]. Puncturing eMBB or mMTC RBs for URLLC users at each mini-slot (with durations of 0.125-0.250 msec) is efficient in terms of reducing the URLLC latency, however, it may degrade eMBB and mMTC Quality of Service (QoS). However, there are some works who considered Dynamic RBS (D-RBS) for their framework [5], [6], [7], [8], but they didn't consider the interference caused by puncturing RBs (or allocating them) for other users in other cells. In other words, they only consider the single cell scenario for their frameworks which is not practical in real applications. Additionally, they only consider the user QoS and the cost of resource usage is neglected, which would impose high cost to the infrastructure if not managed properly [9], therefore a framework should be considered which can jointly support QoS of users and the efficiency of the network resource usage so that it can be implemented in network slicing architecture [10].

To tackle these drawbacks, we **first** propose a novel AI-assisted D-RBS for 5G in multi-cell and multi-agent scenario in which the RG of each cell is partitioned into several Resource Elements (REs) of equal time duration and bandwidth. Based on the service requirements, some REs are aggregated to form one of the three types of RBs: 1) RB type-1 of shape 4×1, i.e., 4 REs in time domain and 1 RE in frequency domain, 2) RB type-2 of shape 1×4, i.e., 1 RE in time domain and 4 REs in frequency domain, and 3) RB type-3 of shape 2×2, i.e., 2 REs in time domain and 2 REs in frequency domain. In other words, the allocated RBs to each user can be constructed using one or more of these REs which could be spanned over specific location in the RG. Note that RB type-1, type-2, and type-3 are used for eMBB, URLLC, and mMTC services, respectively. In our proposed D-RBS, dynamic means that the location of the RBs is not predefined and could be different in RG. Therefore, there are optimal positions for each RB and our objective is to find these optimal positions considering user requirements, network resource usage and the inter-cell interference.

Then, we compare the proposed Intelligent Allocation (IA) scheme with base line approaches, named Random Allocation (RaA) and Over Allocation (OA) in terms of performance and complexity. It is worth noting that we cannot compare our proposed D-RBS method to the existing D-RBS methods since they did not consider the inter-cell interference in their framework. Also, our proposed optimization problem is totally different from the exist D-RBS optimization problems since we consider resource efficiency along with QoS for users. In IA, we consider a Deep Reinforcement Learning (DRL) method to RB assignment and transmit power allocation for eMBB and mMTC users at each time slot in order to satisfy their QoS requirements. Moreover, our aim is to find the proper transmit power and position of RBs to be punctured for the incoming URLLC users in the current mini-slot so that the URLLC packet loss probability and latency constraints are satisfied and the negative impact of these puncturing on the eMBB and mMTC users' performance are minimized. Our general objective is to jointly minimize the RB usage and transmit power. In RaA, after allocating RBs and transmit power to eMBB and mMTC users at each time slot, some of the already allocated eMBB or mMTC RBs are reassigned to the incoming URLLC users with uniform transmit power in the current mini-slot without take into account the negative impact of these puncturing. In OA, we consider extra minimum required rate for eMBB and mMTC users while RaA is performed. In the OA scheme, two weights are considered to hypothetically increase the required rate of eMBB and mMTC users so that if puncturing happens on eMBB RBs or mMTC ones, the minimum required rate for these users remains satisfied. In this approach, our objective is to find the optimum value of the weights by DRL based methods so that by consuming more network resources (RB and transmit power allocation), the average user drop rate decreases remarkably. Most of the RA puncturing problems are NP-hard due to the non-convexity of the objective functions, multiple optimization variables, and highly nonlinear constraints. Thus, it is very difficult to obtain the optimal solution in an acceptable time by traditional solutions, especially in large scale networks. In this paper, we first utilize Single-agent (SA) DRL based methods for dynamic RBs assignment and transmit power allocation. Then, we extend our proposed solution to a Multi-agent (MA) ones, in which the BSs as agents must interact and cooperate with each other to learn and optimize the collaboration policies and solve unstable learning problems.

II. RELATED WORKS

A. STATIC RESOURCE BLOCK STRUCTURE

The RA problem among 5G network's services has been addressed in several works including [11] and [12]. The authors in [11] study the Downlink (DL) dynamic multiplexing of URLLC and eMBB services in a single cell. Thus, they do not consider the inter-cell interference. The goal is to maximize the utility for eMBB traffic based on the three different eMBB loss rate models, named linear, convex, and threshold models with respect to URLLC service requirement constraints. In [12], the authors study the DL resource scheduling with a mini-slot in 5G networks for URLLC services with the aim of achieving a good trade off between spectral efficiency, latency, and reliability for each link and service flow. However, resource scheduling with a mini-slot can cause high signaling overhead due to sending frequently the channel information. The dynamic scheduling for mMTC and URLLC services is studied in [13] for New Radio (NR) where the authors investigate the performance of frameworks that are with/without feedback using RL-based multi-armed bandit approach. In [13], normal feedback transmission on Hybrid Automatic Repeat Request (HARQ) retransmission and one-shot transmission with lower code rate are considered. However, the authors consider the constant code rate and do not focus on link adaptation and a suitable Modulation and Coding Scheme (MCS) selection. In [14], the authors optimize transmit power, bandwidth, and the number of active antennas in a multi-cell scheme to maximize the energy efficiency under QoS constraints of URLLC users including End-to-end (E2E) delay and total packet loss rate. However, they consider the perfect channel gain for all types of users which impedes the real-time implementation of URLLC services. Also, they illustrate that both schemes are able to effectively deploy various URLLC and mMTC services in NR. Recently, there are several researches on RB design in wireless networks [11], [15], [16], [17]. An overview of 5G deployment challenges is provided in [15]. In [16], the authors provide a survey of key techniques to overcome the new requirements and challenges of mMTC applications. In [18], Coordinated Multi-point (CoMP) is used to meet the delay and reliability requirements of the URLLC users. In [17], the authors provide joint power and sub-channel allocation for a sliced 5G network with respect to both the inter-tier and intra-tier interference constraints. In [19], the authors study orthogonal RA for mMTC and eMBB users. The Uplink (UL) multiplexing of URLLC, eMBB, and mMTC services is studied in [20]. In [21], the authors investigate DL transmission design for URLLC services. Joint RA in UL and DL based on effective bandwidth and effective capacity to ensure the QoS for URLLC is considered in [22] and [23]. Joint UL and DL bandwidth optimization with respect to delay constraints to guarantee both packet loss constraints and E2E delay requirement is considered in [23]. In [24], the authors propose a packet delivery mechanism for URLLC. The goal is to reduce the bandwidth required to guarantee the queuing delay based on a statistical multiplexing queuing model. The network slicing based on orthogonal and non-orthogonal radio RA in the UL is considered for the three types of services of 5G in [25]. In [26], the authors optimize power and bandwidth allocation across radio access network slices and users which have heterogeneous QoS requirements. The goal is to maximize both throughput and energy efficiency in the sliced radio access network. In [13] and [27], the authors investigate the dynamic DL RA for eMBB and URLLC services on the same time/frequency resources. The impact of the 5G frame structure on URLLC performance is investigated in [28]. The performance of flexible TTI to support higher traffic load is investigated in [29] and [30]. A 5G frame structure design is considered in [12] to support user's service requirements. Reference [31] explores the multiplexing of eMBB and URLLC services in a wireless-powered communication network. A hybrid access point manages wireless energy transfer and information reception. Preemptive puncturing is utilized to multiplex URLLC traffic onto eMBB transmission. The objective is to jointly allocate subcarriers, time, and energy to maximize the uplink eMBB sum rate while considering URLLC latency, user battery capacity, and subcarrier availability. [32] examines the use of statistical channel knowledge in point-to-point URLLC transmission, exploring various hybrid automatic repeat request schemes and SNR feedback from failed packets to enhance transmission efficiency. The problem is framed as a long-term power minimization issue under URLLC requirements. A DRL agent utilizing proximal policy optimization is employed to dynamically regulate transmit power and coding rate to address the problem. In [33], the authors suggest a method for sub-channel allocation and power control to enable massive access in NOMA-URLLC networks. The problem is modeled as a multi-agent reinforcement learning problem, and a DQN-MARL algorithm is proposed to ensure reliability and latency requirements of URLLC services.

In all the previous works, the authors have considered a static structure for RBs that does not have enough flexibility to respond to different service requirements. Indeed, they do not consider both outage and bit error rate constraints that are critical for new emerging services in 5G. Since the traditional framing, RA, and user association schemes are not flexible enough and do not consider various service types requirements in their optimization problems, they are not appropriate for 5G services.

B. DYNAMIC RESOURCE BLOCK STRUCTURE

In [6] and [8], the authors consider flexible two dimensional resource allocation in order to maximize the overall system energy efficiency and to minimize the adverse impact of puncturing on eMBB users in a single-cell scenario with eMBB and URLLC users. However, these works assume that the BS knows the Channel State Information (CSI) and do not consider multi-cell scenario as well. They also didn't consider the inter-cell interference for RB allocation (puncturning)

which extremely important for real applications [34]. Also, in network slicing arcitectures, the resource efficiency must be considered for the resource allocation mechanisms [35] which is neglected on the existing D-RBS schemes. In our scheme, in accordance on a realistic assumption, the BS should schedule the URLLC users immediately and can not wait to acquire the CSIs. Joint optimization of RA for eMBB and URLLC users based on the flexible frame structure in considered in [7]. However, this work does not utilize puncturing mechanisms to share resources, and instead considers the dedicated RA for URLLC users which can degrade the system performance. In [36], the aim is to maximize the data rate of eMBB users while guaranteeing the latency of URLLC users by user selection and power allocation in a coexistence problem based on a puncturing technique. The authors utilized a Difference of Convex (DC) programming and a Successive Convex Approximation (SCA) to solve the scheduling problem of power allocation and user selection, respectively. In [37], the authors proposed a spectrum partition scheme for maximizing the eMBB data rate and satisfying the URLLC latency and reliability. In [38], the authors proposed a precoding design in order to minimize the transmit power of a Base Station (BS) while satisfying the QoS of eMBB and URLLC users based on imperfect CSI. Joint Transmission (JT) and Orthogonal Transmission (OT) modes were investigated for satisfying the Block Error Rate (BLER) and data rate of URLLC and eMBB, respectively in [39]. To do this, a low-complexity algorithm was proposed to jointly select the group of BSs and their MCSs in order to meet both URLLC and eMBB requirements.

C. MACHINE LEARNING BASED SOLUTIONS

In [40], the authors use Deep-Q-networks (DQNs) to find optimal policies to jointly allocate sub-carriers and transmit power for a vehicle-to-vehicle communication. The DRL approach is used to maximize the energy efficiency by jointly solving sub-channel assignment, transmit power allocation, and user scheduling [41]. Dynamic sub-channel assignment problems are modeled as a Markov Decision Process (MDP) in [42] and [43] or as Partially Observable MDP (POMDP) in [44]. In these works, the authors use different RL approaches to solve their proposed problems. Also, DQN is applied to maximize the network utility, or to minimize the blocking probability of services in [44], [45], and [46]. In [45], the authors apply a multi-agent RL algorithm to their decentralized multi-user system in which a static channel is considered. A comparison between single-agent RL and multi-agent RL is provided in [47]. Note that cooperative multi-agent systems can be used to solve many complex problems, such as control of multiple robots [48] and multi-player games [49]. In these systems, multiple agents collaborate to achieve common goals [50]. As one of the most well known approaches, several agents are independently trained to maximize their reward and treat others as part of the environment [50]. However, due to the changing policies of other agents, the environment is not fully observable to the agents [51]. The approach in [52] provides the coordinated behavior of multiple agents, however, it is not scalable to larger systems because the number of actions increases exponentially with the number of agents. To overcome the challenge of non-Markovian and non-stationary environments during learning, the centralized training of decentralized policies can be considered for efficient training of multiple agents and access additional state information of other agents [53]. In general, RA policies should be designed based on system features, e.g., channel gains and users arrivals, to provide user's satisfaction and QoS requirements. In practice, these features are unknown and generally time-varying. For instance, the number of user arrivals and the value of channel gains can vary over time because mobile users may change their locations over time. Therefore, with limited resources, it is essential to learn how to optimally update the decision policy given the observations of system features. Recently, the DRL-based approach is introduced as a powerful tool to deal with decision-making in dynamic environments [40] and [47]. This technique learns to eventually find the optimal policy in order to enhance long-term performance through interactions with the environments.

III. MAIN CONTRIBUTIONS

Although there are many D-RBS frameworks which are proven to be more efficient compared to S-RBS schemes in terms of QoS of the 5G triple services such as [5], [6], [7], and [8], they didn't consider the multi-cell scenario with inter-cell interference in their formulation which drastically influence on the resource allocation (puncturning) mechanism. In addition, resource efficiency is another important and challenging issue specially in network slicing architectures [35], which is neglected in the exisitng D-RBS structures. In this paper, we present a new dynamic RB structure for multi-cell networks in the 5G resource grid. Unlike previous works focused on single-cell networks, our design addresses the challenge of inter-cell interference. We consider both latency and reliability requirements for eMBB, mMTC, URLLC services. We propose an IA scheme to multiplex these services, ensuring URLLC's stringent latency and reliability constraints while supporting eMBB and mMTC QoS requirements. We compare our scheme with two low complexity schemes, RaA and OA, analyzing the performance trade-off between network resource usage and user satisfaction. We formulate the D-RBS method as three optimization problems, aiming to minimize network resource usage while satisfying eMBB and mMTC QoS requirements and meeting URLLC's latency and reliability constraints. Due to the complexity of the optimization problems, we utilize a DRL method to optimize transmit power allocation and RB assignment in single-agent and multi-agent scenarios. Considering the afformentioned challenges, our contributions are as follows:

- We design a new dynamic RB structure in multi-cell network with inter-cell interference as D-RBS for

5G resource grid so that the RBs for each eMBB, mMTC, and URLLC services are constructed based on their requirements with different shapes that can be located anywhere in the resource grid, in order to adapt to the dynamic nature of the network and satisfy heterogeneous requirements of different users.

- Furthermore, we propose a novel IA scheme as IA for multiplexing eMBB, mMTC, and URLLC users in order to handle URLLC stringent latency and reliability along with supporting eMBB and mMTC QoS requirements and compare this scheme with two Low complexity and overhead schemes, named RaA and OA. Then, we analyze the performance gain of these three schemes which comprehensively indicates the trade-off between network resource usage and user's satisfaction.
- We formulate our proposed D-RBS method with puncturing schemes as three different optimization problems in which the objectives are minimizing the network resource usage along with satisfying the eMBB and mMTC QoS requirements, while the URLLC latency and reliability constraints are satisfied. Since our optimization problems are nonlinear non-convex with multiple discrete and continuous variables, we use DRL method to optimize the transmit power allocation and RBs assignment in single-agent and multi-agent scenarios and show that the proposed framework scales well with a large number of RBs, BSs, and users.
- We further investigate our proposed schemes from the convergence and computational complexity perspectives. Finally, we study the performance of the proposed schemes and compare it with S-RBS baseline approach using simulations for different network parameters.

The remainder of this paper is organized as follows. System model and descriptions regarding 5G services and requirements along with problem formulation and solution algorithms are presented in Section IV. In Section V, we provide a near-optimal solution by a multi-agent approach. In Sections VI and VII, we provide the convergence proof and the computational complexity of our scheme, respectively. Simulation results are provided in Section VIII. And finally, Section IX concludes this work.

TABLE 1. Main parameters and notations.

Parameter	Description
$\mathcal{B}/b/B$	Set/Index/Number of BSs
$\mathcal{K}_b/K_b$	Set/Number of BS b users
$\mathcal{K}_b^{\mathrm{e}}/\mathcal{K}_b^{\mathrm{m}}/\mathcal{K}_b^{\mathrm{u}}$	Set of BS b eMBB, mMTC, and URLLC users
$K_b^{\mathrm{e}}/K_b^{\mathrm{m}}/K_b^{\mathrm{u}}$	Number of BS b eMBB, mMTC, and URLLC users
$\mathcal{K}/k/K$	Set/Index/Number of users
$\mathcal{T}/t/T$	Set/Index/Number of time slots
δ	Duration of each time slot
W	Bandwidth at each cell
χ/ϑ	Time duration/Frequency size of each RE
$\mathcal{I}/i/I$	Set/Index/Number of RE at time domain of each RG
$\mathcal{J}/j/J$	Set/Index/Number of RE at frequency domain of each RG
$\mathcal{F}^{\mathrm{e}}/\mathcal{F}^{\mathrm{m}}/\mathcal{F}^{\mathrm{u}}$	Set of eMBB/mMTC/URLLC RBs at each RG
$F^{\mathrm{e}}/F^{\mathrm{m}}/F^{\mathrm{u}}$	Number of eMBB/mMTC/URLLC RBs at each RG
$\mathcal{F}^{\mathrm{Tot}}/F^{\mathrm{Tot}}$	Set/Number of RBs at each RG
Π^{u}	Possible position for placing URLLC RBs
o_{ij}^{ft}	Binary variable, $o_{ij}^{ft}=1$ if the RB f includes RE (i,j) at time slot t, otherwise $o_{ij}^{ft}=0$.
$o_{ij}^{ft_m}$	Binary variable, $o_{ij}^{ft_m}=1$ if the RB f includes RE (i,j) at mini-slot t_m, otherwise $o_{ij}^{ft}=0$.
ξ_{bk}^{ft}	Binary variable, $\xi_{bk}^{ft_m}\in\{0,1\}$ denote whether RB f is assigned to user k at BS b at mini-slot t_m.
ξ_{bk}^{ft}	Binary variable, $\xi_{bk}^{ft}\in\{0,1\}$ denote whether RB f is assigned to user k at BS b at time slot t.
P_f^{u}	Uniform power allocated to URLLC RB f.
$R_{bk}^{\mathrm{URLLC},ft_m}$	The achievable rate of URLLC user k in RB f at BS b at mini-slot t_m
$R_{bk}^{\mathrm{eMBB/mMTC},ft}$	The achievable rate of eMBB/mMTC user k in RB f at BS b at time slot t
ε_{bk}^{f}	Decoding error probability for user k assigned to BS b on RB f
γ_{bk}^{ft}	SINR for user k on RB f at BS b at time slot t
$\gamma_{bk}^{ft_m}$	SINR for user k on RB f at BS b at mini-slot t_m
$\boldsymbol{g}/\boldsymbol{p}/\boldsymbol{\xi}$	Channel gain/power allocation/and RBs assignment vectors
ψ	Number of symbol in a RB
d_b	Radius of BS b
$p_{bk}^{ft}/p_{bk}^{ft_m}$	Transmit power of BS b to user k on RB f at time slot t and mini-slot t_m
$g_{bk}^{ft_m}$	Channel power gain between BS b and user k on RB f at mini-slot t_m
N_0	single-sided noise PSD
g_{bk}^{ij,ft_m}	Channel power gain between BS b and user k on RE (i,j) of RB f at mini-slot t_m.
κ	transmitting bits from BS b to user k in the short block-length regime

IV. SYSTEM MODEL AND PROBLEM FORMULATION

In this section, we describe the system model in which BSs dynamically allocate the RBs and transmit power to each user. With considering URLLC puncturing scheme, our aim is to jointly minimize the total transmit power of the network and the number of allocated RBs under constraints of different users' requirements. We consider a multi-small cell DL OFDMA network, where there is one macro BS (MBS) and B BSs in B small cells (i.e., a BS at each small cell). The set of BSs is denoted by $\mathcal{B}=\{b_1,b_2,\ldots,b_B\}$ with $|\mathcal{B}|=B$, indexed by b, and $|.|$ denotes the number of elements in a set. In each cell, the BS employs the available radio resources (i.e., total bandwidth and power budget) for serving users located in its circular region with radius d_b. Within each cell coverage, there are three types of randomly distributed users which request different types of services. In BS b, the sets of active users which request eMBB, mMTC, and URLLC services are denoted by $\mathcal{K}_b^{\mathrm{e}}$, $\mathcal{K}_b^{\mathrm{m}}$, and $\mathcal{K}_b^{\mathrm{u}}$, respectively, with $|\mathcal{K}_b^{\mathrm{e}}|=K_b^{\mathrm{e}}$, $|\mathcal{K}_b^{\mathrm{m}}|=K_b^{\mathrm{m}}$ and $|\mathcal{K}_b^{\mathrm{u}}|=K_b^{\mathrm{u}}$. The set of all users in BS b is denoted by $\mathcal{K}_b=\mathcal{K}_b^{\mathrm{e}}\bigcup\mathcal{K}_b^{\mathrm{m}}\bigcup\mathcal{K}_b^{\mathrm{u}}$, with $K_b=|\mathcal{K}_b|$. The set of all users in the proposed system is denoted by $\mathcal{K}=\bigcup_{b\in\mathcal{B}}\mathcal{K}_b$, indexed by k and $K=|\mathcal{K}|$ denotes the total number of users in the proposed system. The main parameters and notations are summarized in Table 1.

A. RESOURCE BLOCK STRUCTURE

We assume that time is slotted into T time slots with equal duration δ, denoted by $\mathcal{T} = \{t_1, t_2, \ldots, t_T\}$, indexed by t. The radio resources are scheduled among eMBB and mMTC users at beginning of each time slot. We assume that the duration of each time slot δ is relatively short, such that mobile users can be considered quasi-static during a single time slot. At each time slot, a time-frequency resource of δ seconds and W Hz is used by each cell. Each time slot is divided into J mini-slots of duration χ for scheduling URLLC users. The radio resources are scheduled among URLLC users at beginning of each mini-slot. In other words, in the proposed frame structure, the time-frequency resource is divided into REs with the time duration χ and frequency size ϑ. Therefore, at each time slot, all the REs could be shown by a matrix $[\boldsymbol{A}]_{I\times J}$ with dimension $J = \frac{\delta}{\chi}$ and $I = \frac{W}{\vartheta}$, where each element of matrix, a_{ij} denotes RE at (i, j) position of time-frequency domain (i.e., RG). The set of REs in each BS at each time slot is denoted by $\mathcal{I} \times \mathcal{J}$ where $\mathcal{I}$ and $\mathcal{J}$ are the sets of REs in frequency and time domains, respectively, with $|\mathcal{I}| = I$, and $|\mathcal{J}| = J$. We consider three types of $\vartheta \times \chi$ shaped RBs: 1) RB type-1 of shape 4×1, 2) RB type-2 of shape 1×4, and 3) RB type-3 of shape 2×2, see Fig. 1(a). Each RB type consists of four adjacent REs. For example, in Fig. 1(b), $I = 8, J = 4$, hence there are $4\times8 = 32$ REs. Let $o_{ij}^{ft} = 1$ if the RB f includes RE (i, j) at time slot t, otherwise $o_{ij}^{ft} = 0$. Placing these types of RBs at all possible positions of the RG, we generate the set of all candidate RBs for RBs type-1, type-2 and type-3 which is denoted by $\mathcal{F}^{\mathrm{u}}$, $\mathcal{F}^{\mathrm{e}}$, and $\mathcal{F}^{\mathrm{m}}$ with $|\mathcal{F}^{\mathrm{u}}| = F^{\mathrm{u}}$, $|\mathcal{F}^{\mathrm{e}}| = F^{\mathrm{e}}$, and $|\mathcal{F}^{\mathrm{m}}| = F^{\mathrm{m}}$, respectively. We assume that the arrival services are assigned to the three types of RBs. The set of possible RBs at each BS (or cell) are denoted by $\mathcal{F}^{\mathrm{Tot}} = \mathcal{F}^{\mathrm{u}} \cup \mathcal{F}^{\mathrm{e}} \cup \mathcal{F}^{\mathrm{m}}$. We assume that RBs of type-2 and 3 can be placed at all possible positions of the RG, while RBs type-1 can be only placed at certain positions, Π^{u}.[1] For example, on RG in Fig. 1 (b), for each mini-slot, there are $\Pi^{\mathrm{u}} = 2$ possible positions for RBs type-1, i.e. first four REs and the last four REs. One or more RBs can be assigned to eMBB or mMTC user at a time slot. Let the binary variables $\xi_{bk}^{ft} \in \{0, 1\}$ denote whether RB f is assigned to user k at BS b at time slot t. In other words, the binary-valued RB-association factor ξ_{bk}^{ft} represents both RB and BS assignments for user k of BS b on RB f at time slot t, i.e., $\xi_{bk}^{ft} = 1$ when BS b allocates RB f to user k, and $\xi_{bk}^{ft} = 0$, otherwise. In order to avoid intra-cell interference, i.e., interference between different eMBB and mMTC RBs, no RE can be assigned to more than one RB. To satisfy that there is no overlap among the chosen REs, the following orthogonality constraint must be met:

$$\sum_{f\in\mathcal{F}^{\mathrm{e}}\cup\mathcal{F}^{\mathrm{m}}} \sum_{k\in\mathcal{K}^{\mathrm{e}}\cup\mathcal{K}^{\mathrm{m}}} o_{ij}^{ft}\xi_{bk}^{ft} \leq 1, \forall b, i, j, t. \tag{1}$$

Therefore, for user k in cell b, the assigned RB is constructed by aggregation of one or more REs. Therefore, all users are multiplexed in an orthogonal fashion to the RBs. Note that the eMBB/mMTC users are allocated in orthogonal fashion which holds with (1) and the punctured RBs of eMBB/mMTC, are orthogonally allocated between URLLC users with the variable $\xi_{bk}^{ft_m}$, which is a binary variable for the puncturing the already allocated RB f at mini-slot t_m for URLLC user k in BS (cell) b. Similarly, to satisfy that there is no overlap among chosen URLLC RBs, the following orthogonality constraint must be met

$$\sum_{f\in\mathcal{F}^{\mathrm{u}}} \sum_{k\in\mathcal{K}^{\mathrm{u}}} o_{ij}^{ft_m}\xi_{bk}^{ft_m} \leq 1, \forall b, i, j, t_m. \tag{2}$$

The URLLC user's packets are randomly generated at each mini-slot. Therefore, the number of URLLC users at each time slot is random.

Accordingly, we decompose the RA problem in two different steps: 1) the first step is the eMBB and mMTC RB and transmit power allocation that is performing at the beginning of each time slot, and 2) the second step is the URLLC puncturing decision at each min-slot in which we consider our proposed scheme and compare it with two low complexity and overhead schemes as RaA and OA. In the following, three schemes named IA scheme, RaA, OA scheme are proposed. In IA scheme, we allocate RBs to eMBB and mMTC users at each time slot and reallocate (puncture) of eMBB or mMTC RBs to URLLC users at each mini-slot. Our aim is to minimize the negative impact of URLLC puncturing on eMBB or mMTC users' data rate and provide the reliability and latency requirement for URLLC users. In other words, after allocation eMBB and mMTC types of RBs to eMBB and mMTC users, respectively, at time-slots, URLLC RBs will be intelligently punctured form allocated eMBB and mMTC RBs at mini-slots of each time-slot so that it satisfies the reliability and latency requirement for URLLC users and has the least negative effect on eMBB or mMTC users' data rate. Although IA scheme provides near-optimal performance for our proposed system, it has complexity and signaling overhead. Therefore, we proposed two low complexity and low signaling overhead schemes named RaA scheme and OA scheme. In RaA scheme, URLLC RBs are randomly punctured from eMBB or mMTC allocated RBs. Although this scheme satisfies low latency constraint for URLLC users, it is not necessarily satisfying the eMBB and mMTC users' requirement. In other words, these puncturing may have negative effect on eMBB or mMTC users' data rate. Therefore, there is a trade-off between the complexity cost and user satisfaction for service providers. To compensate for negative effect of URLLC puncturing on eMBB or mMTC users' data rate in RaA scheme, we proposed another scheme named OA scheme,

[1] For IA, all types of RBs can be placed at all possible position, however for OA and RaA, all types of RBs can be placed at certain positions where RBs do not have any overlap with each others, because in OA and RaA, we do not have any control over RB design.

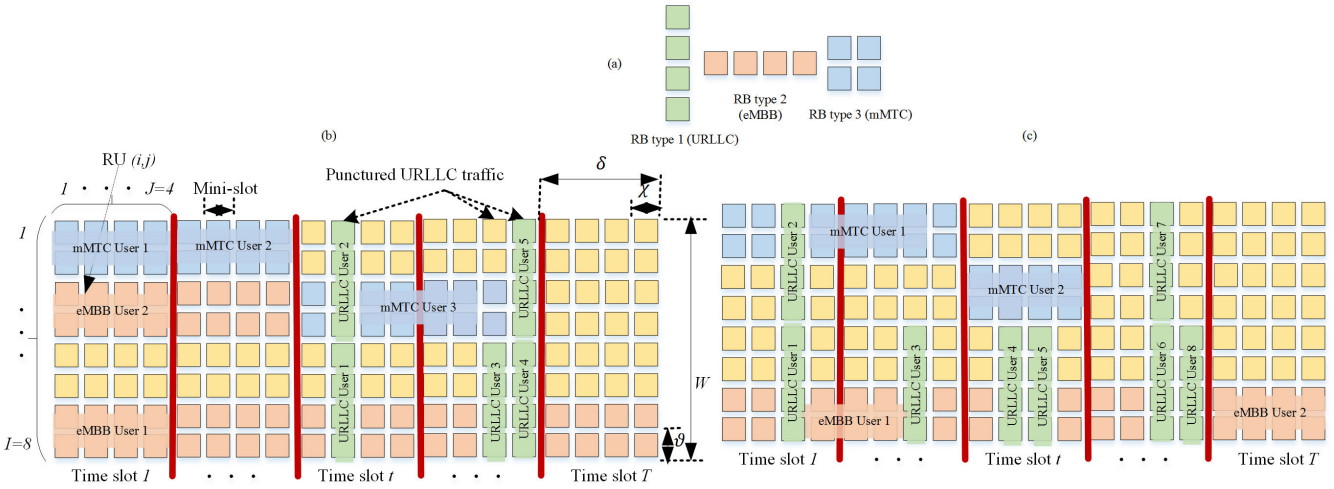


FIGURE 1. a) Three types of RBs structure: 1) RB type-1 of shape 4 × 1. 2) RB type-2 of shape 1 × 4, and 3) RB type-3 of shape 2 × 2, b) Total RBs structure at cell 1: with $I = 8$ and $J = 4$, there are $4 \times 8 = 32$ REs, for each mini-slot, there are $\Pi^{\mathrm{u}} = 2$ possible positions for RBs type-1, i.e. first four REs and the last four REs. c) Total RBs structure at cell 2: with $I = 8$ and $J = 4$, there are $4 \times 8 = 32$ REs.

where more RBs are allocated to eMBB and mMTC users so that if these allocated RBs are punctured by URLLC users, the rates of eMBB and mMTC users do not decrease. This scheme in addition to providing low complexity and signaling overhead, minimizes the negative effect on eMBB or mMTC users' data rate.

B. INTELLIGENT ALLOCATION

In this method, in addition to RA of eMBB and mMTC users at each time slot, we also consider allocation of URLLC RBs and re-allocation of eMBB or mMTC RBs to URLLC users at each mini-slot in order to 1) minimize the total transmit power and the number of allocated RBs, 2) satisfy the requirement of eMBB and mMTC users, 3) minimize negative impact of URLLC puncturing on eMBB or mMTC users data rate, and 4) provide the reliability and latency requirement for URLLC users.

1) URLLC RELIABLE TRANSMISSION AND DATA RATE BASED ON FINITE BLOCK-LENGTH CODING

With carrier and slot aggregation capability, already in use [54], we can aggregate radio carriers (in the same band or across disparate bands) and slots of different small blocks to construct RBs to meet the requirements of users. The achievable rate of URLLC user $k \in \mathcal{K}^{\mathrm{u}}$ in RB $f \in \mathcal{F}^{\mathrm{u}}$ at mini-slot t_m with finite block length can be accurately approximated as follows [55] and [56]:

$$R_{bk}^{\mathrm{URLLC},ft_m}(\varepsilon_{bk}^{f}) \approx \frac{\chi}{\ln 2}\left[4\vartheta \ln\left(1+\gamma_{bk}^{ft_m}(\boldsymbol{p},\boldsymbol{\xi},\boldsymbol{g})\right) - \sqrt{\frac{v_{bk}^{ft_m}}{\psi}}\Theta_{Q}^{-1}(\varepsilon_{bk}^{f})\right], \quad \forall b \in \mathcal{B}, k \in \mathcal{K}^{\mathrm{u}}, f \in \mathcal{F}^{\mathrm{u}}, t_m, \tag{3}$$

where

$$\gamma_{bk}^{ft_m}(\boldsymbol{p},\boldsymbol{\xi},\boldsymbol{g}) = p_{bk}^{ft_m} g_{bk}^{ft_m} \Big/ \Big(\sum_{\bar{b}\in\mathcal{B}\setminus\{b\}} \sum_{\bar{k}\in\mathcal{K}_{\bar{b}}^{\mathrm{u}}} \xi_{\bar{b}\bar{k}}^{ft_m} p_{\bar{b}\bar{k}}^{ft_m} g_{\bar{b}k}^{ft_m} + \sum_{\bar{b}\in\mathcal{B}\setminus\{b\}} \sum_{\bar{k}\in\mathcal{K}_{\bar{b}}^{\mathrm{e}}\cup\mathcal{K}_{\bar{b}}^{\mathrm{m}}} \sum_{\bar{f}\in\mathcal{F}^{\mathrm{e}}\cup\mathcal{F}^{\mathrm{m}}} \sum_{i,j\in f} \xi_{\bar{b}\bar{k}}^{\bar{f}t} o_{ij}^{ft_m} o_{ij}^{\bar{f}t} \frac{p_{\bar{b}\bar{k}}^{\bar{f}t}}{L_f} g_{\bar{b}k}^{ft_m} + 4\vartheta N_0 \Big), \tag{4}$$

in which $\sum_{\bar{b}\in\mathcal{B}\setminus\{b\}} \sum_{\bar{k}\in\mathcal{K}_{\bar{b}}} \xi_{\bar{b}\bar{k}}^{ft_m} p_{\bar{b}\bar{k}}^{ft_m} g_{\bar{b}k}^{ft_m}$ is interference from other BSs that allocate the same URLLC RB f to their serving URLLC users at mini-slot t_m, and $\sum_{\bar{b}\in\mathcal{B}\setminus\{b\}} \sum_{\bar{k}\in\mathcal{K}_{\bar{b}}^{\mathrm{e}}\cup\mathcal{K}_{\bar{b}}^{\mathrm{m}}} \sum_{\bar{f}\in\mathcal{F}^{\mathrm{e}}\cup\mathcal{F}^{\mathrm{m}}} \sum_{i,j\in f} \xi_{\bar{b}\bar{k}}^{\bar{f}t} o_{ij}^{ft_m} o_{ij}^{\bar{f}t} \frac{p_{\bar{b}\bar{k}}^{\bar{f}t}}{L_f} g_{\bar{b}k}^{ft_m}$ is interference form other BSs that allocate eMBB or mMTC RBs to their serving eMBB or mMTC users which have overlap with URLLC RB f. Note that the transmit power of a RB is equally distributed over the frequency domain. Therefore, for example for each URLLC RB which occupies four REs in the frequency domain, to find the power of a RE, the transmit power at each RB, P^{u} is divided by 4. Therefore, $L_f = 4$ if $f \in \mathcal{F}_b^{\mathrm{u}}$, $L_f = 2$ if $f \in \mathcal{F}_b^{\mathrm{m}}$, and $L_f = 1$ otherwise. $\Theta_Q^{-1}(.)$ is the inverse of the Gaussian-Q function, $v_{bk}^{ft_m} = 1 - \frac{1}{(1+\gamma_{bk}^{ft_m}(\boldsymbol{p},\boldsymbol{\xi},\boldsymbol{g}))^2}$, and ε_{bk}^{f} denotes the decoding error probability for user k assigned to BS b on RB f at mini-slot t_m. $\boldsymbol{g} = [g_{11}^{11}, \ldots, g_{bk}^{ft_m}, \ldots, g_{bk}^{ft}, \ldots, g_{BK}^{FT}]^T$, $\boldsymbol{p} = [p_{11}^{11}, \ldots, p_{bk}^{ft_m}, \ldots, p_{bk}^{ft}, \ldots, p_{BK}^{FT}]^T$, and $\boldsymbol{\xi} = [\xi_{11}^{11}, \ldots, \xi_{bk}^{ft_m}, \ldots, \xi_{bk}^{ft}, \ldots, \xi_{BK}^{FT}]^T$ denote the channel gain, the power allocation, and RBs assignment vectors. p_{bk}^{ft} and $p_{bk}^{ft_m}$ are the transmit power of BS b to user k on RB f at time slot t and mini-slot t_m, respectively, $g_{bk}^{ft_m}$ is the channel power gain between BS b and user k on RB f at mini-slot t_m, and N_0 is the single-sided noise power-spectral-density (PSD). For each RB, the channel gain depends on the position of RB and its corresponding REs, and is determined by averaging over gains of its REs as $g_{bk}^{ft_m} = \sum_{i,j\in f} o_{ij}^{ft_m} g_{bk}^{ij,ft_m}/4$, where

g_{bk}^{ij,ft_m} is the channel power gain between BS b and user k on RE (i,j) of RB f at mini-slot t_m.

The number of symbols in the block is $\psi = \chi\vartheta$. When transmitting κ bits from BS b to user k in the short block-length regime, by setting $\chi R_{bk}^{\mathrm{URLLC},ft_m} = \kappa$, the decoding error probability over Rayleigh fading channel can be obtained from (3) as follows:

$$\varepsilon_{bk}^{f} \approx \mathbb{E}_{\boldsymbol{g}}\left\{\Theta_Q\left(\sqrt{\frac{\vartheta}{v_{bk}^{ft_m}}}\left[\ln\left(1+\gamma_{bk}^{ft_m}(\boldsymbol{p},\boldsymbol{\xi},\boldsymbol{g})\right)-\frac{\kappa\ln 2}{\chi\vartheta}\right]\right)\right\}, \quad \forall b\in\mathcal{B}, k\in\mathcal{K}^{\mathrm{u}}, f\in\mathcal{F}^{\mathrm{u}}, \tag{5}$$

where $\mathbb{E}(x)$ denotes the expected value of x.

The following constraint should be satisfied to ensure that the packet loss probability of user k in RB f is equal or below the threshold value $\varepsilon_{bk}^{\max,f}$ at mini-slot t_m:

$$\mathbb{E}_{\boldsymbol{g}}\left\{\Gamma(\gamma_{bk}^{ft_m}(\boldsymbol{p},\boldsymbol{\xi},\boldsymbol{g})\right\} \le \varepsilon_{bk}^{\max,f}, \forall t_m, b\in\mathcal{B}, k\in\mathcal{K}_b^{\mathrm{u}}. \tag{6}$$

Note that the average is taken at each time slot over the small-scale channel gains conditioned on large-scale channel gains. It is also worth noting that we assume we have the probability density function (pdf) of the URLLC channel gains and based on the given large-scale channel gains, the decoding error probabilities is only depend on small-scale channel fading [25], [57].

2) eMBB AND mMTC DATA RATE BASED ON SHANNON CAPACITY MODEL

eMBB and mMTC data rates affect by puncturing URLLC traffics. In IA method, we perform resource allocation for URLLC users at each mini-slot, hence we know the position of URLLC puncturing by reallocating them the eMBB or mMTC RBs. Therefore, the eMBB or mMTC instantaneous transmission rate between BS b and user k on RB f at time slot t is defined as:

$$R_{bk}^{\mathrm{eMBB/mMTC},ft}(\boldsymbol{p},\boldsymbol{\xi},\boldsymbol{g}) = \sum_{j=1}^{Z^t} 4\chi\vartheta\left(1-\frac{\sum_{\bar{k}\in\mathcal{K}_b^{\mathrm{u}}}\sum_{i,j\in f}\xi_{b\bar{k}}^{ft}o_{ij}^{ft}}{Z^t}\right) \times \log_2\left(1+\frac{\sum_{f\in\mathcal{F}}\sum_{k\in\mathcal{K}}\xi_{bk}^{ft}p_{bk}^{ft}g_{bk}^{ft}}{I_{bk}^{ft}(\boldsymbol{p},\boldsymbol{\xi},\boldsymbol{g})+\vartheta N_0}\right), \tag{7}$$

where $\sum_{\bar{k}\in\mathcal{K}_b^{\mathrm{u}}}\sum_{i,j\in f}\xi_{b\bar{k}}^{ft}o_{ij}^{ft}$ is the number of punctured mini-slots from the RB f of user k at time slot t and $I_{bk}^{ft}(\boldsymbol{p},\boldsymbol{\xi},\boldsymbol{g})$ denotes interference form other cells on RB f at time slot t which can be expressed as:

$$I_{bk}^{ft}(\boldsymbol{p},\boldsymbol{\xi},\boldsymbol{g}) = \sum_{\bar{b}\in\mathcal{B}\setminus\{b\}}\sum_{\bar{k}\in\mathcal{K}_{\bar{u}}}\sum_{i,j\in f}\xi_{\bar{b}\bar{k}}^{ft_m}o_{ij}^{ft_m}\frac{p_{\bar{b}\bar{k}}^{ft_m}}{L_f}g_{\bar{b}k}^{ij,ft_m} + \sum_{\bar{b}\in\mathcal{B}\setminus\{b\}}\sum_{\bar{k}\in\bar{\mathcal{K}}}\sum_{i,j\in f}\xi_{\bar{b}\bar{k}}^{ft}o_{ij}^{ft}\frac{p_{\bar{b}\bar{k}}^{ft}}{L_f}g_{\bar{b}k}^{ij,ft}, \tag{8}$$

where the first term denotes the interference form URLLC users of other cells which have overlap with RB f and second term denotes the interference form eMBB or mMTC users of other cells which have overlap with RB f. For user k which requests eMBB service, the following constraint should be satisfied to ensure that the average data rate of eMBB user $k\in\mathcal{K}_b^e$ is equal or above the required minimum data rate:

$$\mathbb{E}_{\boldsymbol{g}}\left\{\frac{1}{T}\sum_{t\in\mathcal{T}}\sum_{f\in\mathcal{F}}\xi_{bk}^{ft}R_{bk}^{ft}(\boldsymbol{p},\boldsymbol{\xi},\boldsymbol{g})\right\} \ge R_{bk}^{\min,\mathrm{e}}, \forall b\in\mathcal{B}, k\in\mathcal{K}_b^{\mathrm{e}}, \tag{9}$$

where $R_{bk}^{\min,\mathrm{e}}$ denotes the required data rate of eMBB user k at BS b. The mMTC users require fixed, typically low, transmission rate and PER on the order of 10^{-1} [1]. However, to guarantee the mMTC users data rate requirements, the following constraint should be applied:

$$\sum_{f\in\mathcal{F}}\sum_{t\in\mathcal{T}}\xi_{bk}^{ft}R_{bk}^{ft}(\boldsymbol{p},\boldsymbol{\xi},\boldsymbol{g}) \ge R_{bk}^{\min,\mathrm{m}}, \forall b\in\mathcal{B}, k\in\mathcal{K}_b^{\mathrm{m}}. \tag{10}$$

3) UBIQUITOUS CONNECTIVITY FOR mMTC USERS

In order to satisfy ubiquitous connectivity for mMTC devices, we aim to maximize the number of allocated RBs for mMTC users. Let c_{m_k} be a binary indicator variable to define whether the mMTC user k is connected to the BS or not, where $c_{m_k}=1$ if the data rate of mMTC user k is equal or greater than the minimum data rate $R_{bk}^{\min,\mathrm{m}}$. In order to satisfy the QoS of mMTC service, the following constraint must be met:

$$\sum_{f\in\mathcal{F}^{\mathrm{m}}}\sum_{k\in\mathcal{K}_b^{\mathrm{m}}}\sum_{t\in\mathcal{T}}\xi_{bk}^{ft}c_{m_k} \ge C_M^{\mathrm{th}}, \ \forall b\in\mathcal{B}, \tag{11}$$

where C_M^{th} is the threshold value of the minimum number of connected mMTC devices.

4) PROBLEM FORMULATION

For IA scheme, we aim to jointly minimize the total transmit power and the number of allocated RBs along with satisfying the URLLC reliability an latency constraint which can be formulated as follows:

$$\min_{\boldsymbol{p},\boldsymbol{\xi}} \ \lim_{T\to\infty}\frac{1}{T}\sum_{t\in\mathcal{T}}\sum_{k\in\mathcal{K}}\sum_{f\in\mathcal{F}}\sum_{b\in\mathcal{B}}\left(\varrho_2^i\xi_{bk}^{ft}p_{bk}^{ft}+\varrho_3^i\xi_{bk}^{ft}\right), \tag{12a}$$

$$\text{s.t. } p_{bk}^{ft}\ge 0, \xi_{bk}^{ft}\in\{0,1\}, \forall f\in\mathcal{F}, t\in\mathcal{T}, b\in\mathcal{B}, k\in\mathcal{K}, \quad (1),(2),(6),(9),(10),(11), \tag{12b}$$

where ϱ_2^i and ϱ_3^i are the weights which act as balancing parameters of our objective function. It is worth mentioning that equation (6) is for ensuring the reliability of the URLLC users. Unlike RaA and OA schemes, where upon arriving the URLLC packets at each mini-slot, the BSs randomly punctures some of the already allocated RBs to the arrived URLLC services. On the other hand, in the IA optimization problem, the BSs intelligently puncture the already allocated RBs to the URLLC services to satisfy the eMBB, mMTC, and URLLC constraints, i.e. (1), (2), (6), (9), (10), and (11). We do

not consider constraint (6) for the RaA and OA problems to decrease the complexity of the optimization problem as well as decrease the signaling overhead of the whole system.

C. LOW COMPLEXITY AND OVERHEAD SCHEMES

1) RANDOM ALLOCATION

In addition, the URLLC service requires low-latency and very high reliability transmission with packet loss probability lower than 10^{-7} [1]. This imposes transmission of each URLLC user without waiting to acquire the CSI and its RB should be spanned only on a single mini-slot but can have multiple frequencies. The CSI acquisition requires some time and leads to a significant increase in the latency of URLLC users. Therefore, due to the lack of CSI, no transmit power or rate adaptation is possible for the URLLC users and the BSs should allocate the transmit power and RBs among the URLLC users without knowing the CSI. Hence, we consider the equal power allocation for each URLLC RB, P_f^{u}, $\forall f \in \Pi_f^{\mathrm{u}}$. We assume that the data rate of all URLLC users is relatively low which can be guaranteed by one URLLC RB. Therefore, at each mini-slot, one 1×4 shape RB is randomly punctured from eMBB or mMTC allocated RBs and will be reallocated to the arrived URLLC user with constant transmit power equal to P_f^{u}. Although with this puncturing method, we do not have any control for eMBB and mMTC service constraints, its complexity cost is low and acceptable and the latency constraint for URLLC users are satisfied. In other words, there is a trade-off between the complexity cost and user satisfaction for service providers. We model the number of URLLC users arriving per mini-slot j of time slot t at BS b by a Poisson process with rate $\bar{\lambda}$. In the majority of prior literature, specifically [58], [59], the modeling of delay-sensitive wireless multimedia and vehicle control in scenarios such as V2X has been conducted. In these works, the time-critical traffic is not periodic, and thus, the authors have employed a Poisson process to model the behavior. This process involves URLLC users either transmitting over the RBs allocated to eMBB users or punching them. The Poisson process is a widely used model for traffic modeling in various communication systems, including URLLC. The rationale behind utilizing the Poisson process for modeling URLLC traffic is that it assumes events occur independently of each other. This assumption is reasonable for many types of traffic, such as packet arrivals in a network [60], where the arrival of one packet does not significantly impact the arrival of others. Despite the assumption of independence, the Poisson process can still capture bursty behavior [61]. The process allows for variations in the arrival rate over time, which can reflect the burstiness observed in real-world traffic. We assume that there are Π^{u} URLLC RBs for each mini-slot of each time slot. For each mini-slot j at time slot t, the probability of URLLC user assignment to RB $f \in \Pi^{\mathrm{u}}$ is calculated as follows:

$$\mathcal{P}_{bj}^{ft} = \sum_{k=1}^{\Pi^{\mathrm{u}}-1} \mathcal{P}\{X_{bjt}^{\mathrm{u}} = k\} \frac{\binom{\Pi^{\mathrm{u}}-1}{k-1}}{\binom{\Pi^{\mathrm{u}}}{k}} + \sum_{k=\Pi^{\mathrm{u}}}^{\infty} \mathcal{P}\{X_{bjt}^{\mathrm{u}} = k\}, \quad \forall b, j, t, f \in \mathcal{F}^{\mathrm{u}}, \tag{13}$$

where $\mathcal{P}\{X_{bjt}^{\mathrm{u}} = k\} = \frac{\bar{\lambda}^k e^{-k}}{k!}, \forall k \in \mathcal{K}^{\mathrm{u}}$ is used for finding the probability of a number of arriving URLLC users in a mini-slot.[2]

For eMBB and mMTC users, the instantaneous transmission rate between the b^{th} BS and the k^{th} user on RB f at time slot t is defined as

$$\begin{aligned}\tilde{R}_{bk}^{\mathrm{eMBB/mMTC},ft}(\boldsymbol{p},\boldsymbol{\xi},\boldsymbol{g}) &= \sum_{z_t^{\mathrm{u}}=1}^{Z^t} 4\chi\vartheta\mathcal{P}\{Y_{bt}^{\mathrm{u}} = z_t^{\mathrm{u}}\}\left(1-\frac{z_t^{\mathrm{u}}}{Z^t}\right) \\ &\quad \times \log_2\left(1+\frac{\sum_{f\in\mathcal{F}}\sum_{k\in\mathcal{K}}\xi_{bk}^{ft}p_{bk}^{ft}g_{bk}^{ft}}{\tilde{I}_{bk}^{ft}(\boldsymbol{p},\boldsymbol{\xi},\boldsymbol{g})+\vartheta N_0}\right),\end{aligned} \tag{14}$$

where z_t^{u} is the number of punctured mini-slots at time slot t, Z^t is the number of mini-slots in time slot t. $\mathcal{P}\{Y_{bt}^{\mathrm{u}} = z_t^{\mathrm{u}}\}$ denotes the probability that the number of punctured mini-slots is z_t^{u} in a time slot which can be calculated as

$$\mathcal{P}\{Y_{bt}^{\mathrm{u}} = z_t^{\mathrm{u}}\} = \binom{Z^t}{z_t^{\mathrm{u}}}(\mathcal{P}_{bj}^{ft})^{z_t^{\mathrm{u}}}(1-\mathcal{P}_{bj}^{ft})^{Z^t-z_t^{\mathrm{u}}}, \tag{15}$$

where $\tilde{I}_{bk}^{ft}(\boldsymbol{p},\boldsymbol{\xi},\boldsymbol{g})$ denotes the interference from other cells on RB f at time slot t over user k at the b^{th} BS which can be expressed as

$$\begin{aligned}\tilde{I}_{bk}^{ft}(\boldsymbol{p},\boldsymbol{\xi},\boldsymbol{g}) &= \sum_{\bar{b}\in\mathcal{B}\setminus\{b\}}\sum_{\bar{k}\in\bar{\mathcal{K}}}\sum_{i,j\in f}\mathcal{P}_{bj}^{ft}\xi_{\bar{b}\bar{k}}^{ft}o_{ij}^{ft}\frac{\boldsymbol{P}^{\mathrm{u}}}{L_f}g_{\bar{b}\bar{k}}^{ij,ft} \\ &\quad + \left(1-\mathcal{P}_{bj}^{ft}\right)\xi_{\bar{b}\bar{k}}^{ft}o_{ij}^{ft}\frac{p_{\bar{b}\bar{k}}^{ft}}{L_f}g_{\bar{b}\bar{k}}^{ij,ft}.\end{aligned} \tag{16}$$

In (16), in the first term, the probability of URLLC user assignment to RE (i,j) is multiplied by the transmit power of URLLC RE and channel gain.[3] For eMBB and mMTC users, the following constraint should be satisfied:

$$\mathbb{E}_{\boldsymbol{g}}\left\{\frac{1}{T}\sum_{f\in\mathcal{F}}\sum_{t\in\mathcal{T}}\xi_{bk}^{ft}\tilde{R}_{bk}^{\mathrm{eMBB/mMTC},ft}(\boldsymbol{p},\boldsymbol{\xi},\boldsymbol{g})\right\} \geq R_{bk}^{\min,\mathrm{e}}, \tag{17}$$

$$\sum_{f\in\mathcal{F}}\sum_{t\in\mathcal{T}}\xi_{bk}^{ft}\tilde{R}_{bk}^{\mathrm{eMBB/mMTC},ft}(\boldsymbol{p},\boldsymbol{\xi},\boldsymbol{g}) \geq R_{bk}^{\min,\mathrm{m}}. \tag{18}$$

We aim to jointly minimize the total transmit power and the number of allocated RBs. Therefore, for RaA scheme, the optimization problem can be formulated as follows:

$$\min_{\boldsymbol{p},\boldsymbol{\xi}}\lim_{T\to\infty}\frac{1}{T}\sum_{t\in\mathcal{T}}\sum_{k\in\mathcal{K}^{\mathrm{e}}\cup\mathcal{K}^{\mathrm{m}}}\sum_{f\in\mathcal{F}^{\mathrm{e}}\cup\mathcal{F}^{\mathrm{m}}}\sum_{b\in\mathcal{B}}\left(\varrho_2^r\xi_{bk}^{ft}p_{bk}^{ft}+\varrho_3^r\xi_{bk}^{ft}\right), \tag{19a}$$

$$\text{s.t. } p_{bk}^{ft} \geq 0, \xi_{bk}^{ft} \in \{0,1\}, \quad (1), (11), (17), (18), \tag{19b}$$

[2] We assume that the statistic of URLLC arrival changes very slow such that we can consider this static over a block of time slots. Additionally, note that the arrival of URLLC users at each mini-slot is assumed to be an i.i.d process and Poisson distributed.

[3] The probability of URLLC user assignment to RE (i,j) belongs to RB f is equal to the probability of URLLC user assignment to RB f.

where ϱ_2^r and ϱ_3^r are the weights which act as balancing parameters of our objective function.

2) OVERALLOCATION FOR eMBB AND mMTC USERS

We consider coefficients $c_k^e \geq 1$ and $c_k^m \geq 1$ for overestimating the minimum required rate of eMBB and mMTC users, while for incoming URLLC users at each mini-slot, like the previous method, an RB from allocated eMBB or mMTC users is randomly punctured. In this method, we imagine that there is no URLLC users to puncture eMBB or mMTC RBs and by this assumption, we reconstruct the constraints (17) and (18) as follows:

$$\sum_{f\in\mathcal{F}}\sum_{t\in\mathcal{T}} \xi_{bk}^{ft}\hat{R}_{bk}^{\text{eMBB/mMTC},ft}(\boldsymbol{p},\boldsymbol{\xi},\boldsymbol{g}) \geq c_k^e R_{bk}^{\min,\text{e}}, \quad \forall b\in\mathcal{B}, k\in\mathcal{K}_b^{\text{e}}, \tag{20}$$

$$\sum_{f\in\mathcal{F}}\sum_{t\in\mathcal{T}} \xi_{bk}^{ft}\hat{R}_{bk}^{\text{eMBB/mMTC},ft}(\boldsymbol{p},\boldsymbol{\xi},\boldsymbol{g}) \geq c_k^m R_{bk}^{\min,\text{m}}, \quad \forall b\in\mathcal{B}, k\in\mathcal{K}_b^{\text{m}}, \tag{21}$$

where

$$\hat{R}_{bk}^{\text{eMBB/mMTC},ft}(\boldsymbol{p},\boldsymbol{\xi},\boldsymbol{g}) = 4\chi\vartheta\log_2\left(1+\frac{\sum_{f\in\mathcal{F}}\sum_{k\in\mathcal{K}}\xi_{bk}^{ft}p_{bk}^{ft}g_{bk}^{ft}}{\hat{I}_{bk}^{ft}(\boldsymbol{p},\boldsymbol{\xi},\boldsymbol{g})+\vartheta N_0}\right), \quad \forall b\in\mathcal{B}, k\in\mathcal{K}_b^{\text{e}}\cup\mathcal{K}_b^{\text{m}}, f\in\mathcal{F}_b^{\text{e}}\cup\mathcal{F}_b^{\text{m}}, t, \tag{22}$$

where $\hat{I}_{bk}^{ft}(\boldsymbol{p},\boldsymbol{\xi},\boldsymbol{g})$ can be expressed as

$$\hat{I}_{bk}^{ft}(\boldsymbol{p},\boldsymbol{\xi},\boldsymbol{g}) = \sum_{\bar{b}\in\mathcal{B}\setminus\{b\}}\sum_{\bar{k}\in\mathcal{K}_{\bar{b}}^{\text{e}}\cup\mathcal{K}_{\bar{b}}^{\text{m}}}\sum_{i,j\in f}\sum_{\bar{f}\in\mathcal{F}_{\bar{b}}^{\text{e}}\cup\mathcal{F}_{\bar{b}}^{\text{m}}} \xi_{\bar{b}\bar{k}}^{\bar{f}t}o_{ij}^{\bar{f}t}\frac{p_{\bar{b}\bar{k}}^{\bar{f}t}}{L_{\bar{f}}}g_{\bar{b}k}^{ij,\bar{f}t}. \tag{23}$$

We set the coefficients in a way so that the imaginary data rate required by the eMBB and mMTC users is considered to be high enough so that by puncturing some of their RBs, the required data rate by eMBB and mMTC users will be provided. We aim to find the minimum value for the coefficients. For OA scheme, we aim to jointly minimize the total transmit power, the number of allocated RBs, and two defined coefficients which can be formulated as follows:

$$\min_{\boldsymbol{p},\boldsymbol{\xi},\boldsymbol{c}} \sum_{k\in\mathcal{K}}\left(\varrho_1 c_k^e + \varrho_1' c_k^m\right) + \lim_{T\to\infty}\frac{1}{T}\sum_{t\in\mathcal{T}}\sum_{k\in\mathcal{K}}\sum_{f\in\mathcal{F}}\sum_{b\in\mathcal{B}}\left(\varrho_2^o\xi_{bk}^{ft}p_{bk}^{ft}+\varrho_3^o\xi_{bk}^{ft}\right), \tag{24a}$$

$$\text{s.t. } p_{bk}^{ft}\geq 0, \xi_{bk}^{ft}\in\{0,1\}, c_k^e\geq 1, c_k^m\geq 1, \quad \forall f\in\mathcal{F}, t\in\mathcal{T}, b\in\mathcal{B}, k\in\mathcal{K}, \quad (1), (11), (20), (21), \tag{24b}$$

where ϱ_1, ϱ_1', ϱ_2^o, and ϱ_3^o are the weights which acts as balancing parameters. In this method we formulate our problem under two assumptions; 1) there is no URLLC users that puncture the eMBB and mMTC RBS, and 2) the eMBB and mMTC users need more data rate than their real requirements. So whenever the RBs of eMBB and mMTC are randomly puncture for the arrived URLLC users, the minimum required data rate of the eMBB and mMTC users are satisfied. The objective in this method is to find the optimum value for the over estimation coefficients so that the network resource usage is minimized.

V. REINFORCEMENT LEARNING BASED SOLUTION

An MDP is a mathematical framework used to model decision-making problems in a stochastic (probabilistic) environment. It consists of a set of states, actions, transition probabilities, and rewards. MDPs are widely used in the field of reinforcement learning and decision theory [62], [63]. Hence, we formulate the power allocation and RB assignment for multiple users as an MDP. In this framework, BSs act as agents, making decisions on power allocation for the users who are in a known state. These decisions lead to a transition in the environment involving both the users and BSs, and in response, the agents receive feedback in the form of rewards. Accordingly, we propose a single-agent reinforcement learning method to solve our problem. Then, we extend our solution to multiple agents. Fig. 2 depicts our proposed single and multi-agent solution algorithms.

A. SINGLE-AGENT (CENTRALIZED)

We consider the BSs and users with their channel gains as the environment and consider a global controller that manages all BSs as shown in the left picture in Fig. 2. We define the state, action, and the reward function for IA, OA, RaA schemes as follows:

- **State space:** The state space at each time slot t is the set of observation of the Macro Base Station (MBS) from the environment. For the RaA and OA schemes, It consists of the channel power gains of the eMBB and mMTC users and the number of active eMBB and mMTC users as $\mathcal{S}_r^t = \mathcal{S}_o^t = \left[\{\boldsymbol{g}^t\}, K_{e,t}, K^{m,t}\right]$, where $\boldsymbol{g}^t$ is the vector of all user channel power gains and $K^{x,t}$, $x\in\{e,m\}$ is the total number of active eMBB/mMTC users. For our proposed IA scheme, in addition to the RaA state space, it consists of the arrived URLLC users, so we define the state space of our proposed OA scheme as $\mathcal{S}_i^t = \left[\{\boldsymbol{g}^t\}, K_{e,t}, K^{m,t}, K^{u,t}\right]$. Note that all the values in the state space are continuous variables.
- **Action space:** The action space for RaA scheme at each time slot t consists of the transmit power and RB allocation to the eMBB and mMTC users as $\mathcal{A}_r^t = \{\xi_k^{ft}, p_k^{ft} | \xi_k^{ft}\in\{0,1\}, 0\leq p_k^{ft}\leq P^{\max}\}$. For the OA scheme, in addition to RaA action space, it consists of the overestimation coefficients and can be defined as $\mathcal{A}_o^t = \{c_k^e, c_k^m, \xi_k^{ft}, p_k^{ft} | \xi_k^{ft}\in\{0,1\}, c_k^e, c_k^m\geq 1, 0\leq p_k^{ft}\leq P^{\max}\}$. Finally, for our IA scheme, in addition to the actions of RaA schemes, it consists of URLLC

transmit power and RB allocation and can be defined as $\mathcal{A}_i^t = \{\xi_k^{ft}, \xi_k^{ft}, p_k^{ft}, p_k^{u,ft} | \xi_k^{ft}, \xi_k^{ft} \in \{0,1\}, 0 \leq p_k^{ft}, p_k^{u,ft} \leq P^{\max}\}$. It is worth noting that the action space of all schemes consist of both continuous and discrete variables.

- **Reward function:** The reward function must be in a way that can satisfy the objective of the problem as well as supporting the respective constraints. We define the reward, $r_{\varkappa}^t$ ($\forall \varkappa \in \{r, o, i\}$) as the objective functions of the three IA, RaA, and OA optimization problems:

$$\text{IA: } r_i^t(s^t, a^t) = \begin{cases} -\sum_{f\in\mathcal{F}}\sum_{b\in\mathcal{B}}\sum_{k\in\mathcal{K}}\left(\varrho_2^i \xi_{bk}^{ft} p_{bk}^{ft} + \varrho_3^i \xi_{bk}^{ft}\right) \\ s.t.\ (1), (6), (17), (18), (2), \\ -\infty \qquad \text{Otherwise.} \end{cases} \tag{25}$$

$$\text{RaA: } r_r^t(s^t, a^t) = \begin{cases} -\sum_{f\in\mathcal{F}}\sum_{b\in\mathcal{B}}\sum_{k\in\mathcal{K}}\left(\varrho_2^r \xi_{bk}^{ft} p_{bk}^{ft} + \varrho_3^r \xi_{bk}^{ft}\right) \\ s.t.(1), (17), (18), \\ -\infty \qquad \text{Otherwise.} \end{cases} \tag{26}$$

$$\text{OA: } r_o^t(s^t, a^t) = \begin{cases} -\sum_{k\in\mathcal{K}}\left(\varrho_1 c_k^e + \varrho_1' c_k^m\right) - \\ \sum_{f\in\mathcal{F}}\sum_{b\in\mathcal{B}}\sum_{k\in\mathcal{K}}\left(\varrho_2^o \xi_{bk}^{ft} p_{bk}^{ft} + \varrho_3^o \xi_{bk}^{ft}\right) \\ s.t.(1), (20), (21). \\ -\infty \qquad \text{Otherwise.} \end{cases} \tag{27}$$

B. MULTI-AGENT (DISTRIBUTED)

We consider each BS as an agent which only controls its corresponding users in the coverage area as shown in the right picture in Fig. 2. The agents have back haul communication with the global controller (MBS) for their training. The difference between the single-agent and multi-agent here is, in multi-agent, each agent chooses its own action individually and the trained by MBS. However, in single-agent method, MBS itself chooses all actions and the training performs by itself. So, in single-agent method, we have centralized action selection and centralized training. Conversely, in multi-agent method we have decentralized action selection and centralized training. Since there is no prior information about the time-varying channel conditions and multiple users which want to access the network simultaneously, the state transition function is unspecified and the optimization problem cannot be solved directly. To overcome this challenge, the reinforcement learning methods, such as Q-learning can be utilized, which require no prior information about the state transition function. Hence, we reformulate our optimization problem using the concept of Q-function. At each decision epoch t, for state s^t, after selecting an action a^t based on the decision policy $\pi : s \rightarrow a$, the agent updates an action-value function (e.g., a Q-table) as follows [64]:

$$Q(s^t, a^t) = (1-\alpha)Q(s^t, a^t) + \alpha(r^{t+1} + \gamma \max_{a'\in\mathcal{A}} Q(s^{t+1}, a')), \tag{28}$$

where $Q(s^t, a^t)$ denotes the expected value for state-action pair (s^t, a^t) at epoch t, γ and α are the reward decay and the learning rate over the interval [0, 1], respectively. The agent can select an action randomly or based on the action decision policy such as ϵ-greedy policy. The optimal policy can be obtained as [64]:

$$(\pi^t)^* = \arg\max_{a^t} Q(s^t, a^t). \tag{29}$$

The agent selects the action corresponding to the maximum Q-value with the probability of $1-\epsilon$ or selects the action randomly with the probability of ϵ, where ϵ is equal to 1 at the beginning of learning and decays with each iterations till it gets to zero. The classic reinforcement learning methods such as Q-learning can be used for obtaining the optimal Q-function value for discrete and small state and action spaces. However, for large state and action spaces, the Deep Neural Networks (DNNs) are used as function approximators to estimate the Q-function. On the other hand, our problem consists of power allocation and RB assignment, which needs continuous and discrete actions and the classical DRL methods, such as Actor Critic (AC) and DQN can not be used to obtain the optimal Q-function value and it takes long time for the Q-function to be converged. Therefore, we utilize the Compound Action Actor Critic (CA2C) method proposed in [65] which consists of both the AC and the DQN methods to handle continuous and discrete actions. The environment of our problem consists of multiple BSs with their corresponding users which have diverse service requests, so considering the single-agent DRL methods cause to increase the signaling overhead. On the other hand, a single-agent has limited capacity to solve the problem with large state and action spaces. Accordingly, cooperation of multiple single-agents to solve a common problem leads to jointly decrease the signaling overhead and increase the capacity of learning to find the near-optimal solution. Therefore, we use multi-agent DRL as a solution for solving our proposed resource allocation problem. In our multi-agent DRL method, each BS acts as an agent and interacts with the environment and takes its own action based on its observation independently of each other. Since the BSs objective is common, they are trained by a centralized critic network. In this approach, agents do not need to acquire global information which can reduce the signaling overhead [66]. Regarding to cooperative multi-agent Deep Deterministic Policy Gradient (MADDPG) introduced in [66], we consider CA2C instead of DDPG for each agent and proposed the Multi-Agent-Compound Action Actor Critic (MA-CA2C) method. The details of the MA-CA2C algorithm are described in Algorithm 1. In Lines 1-3, all variables including actor and critic parameters along with

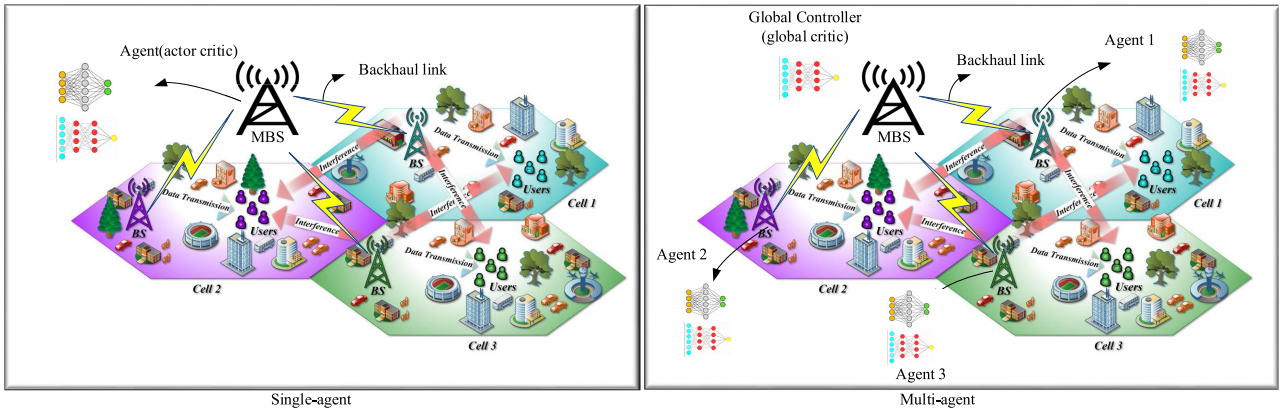


FIGURE 2. **Our proposed system model considering: the single-agent (left), where both training and execution (i.e, resource allocation decisions) stages are taken at single agent (centralized agent) consisting of both actor and critic networks located at MBS, and multi-agent (right) schemes, where only training stage is performed centrally at a critic network located MBS and execution (i.e, resource allocation decisions) stages are distributively taken at each actor network located at each BS.**

the replay buffer size are initialized. The most outer loop (Line 4) iterates based on the number of episodes. The middle loop (Line 5) is for each training step, which defined here as cycles. And the most inner loop (Line 6) iterates based on the number of agents (BSs). From Lines 7-12, each agent interacts with environment by performing action, receiving reward, and saving these transitions into their experience replay D_b. In Lines 13 and 14, the training process is performed. And finally, the neural network parameters are updated in Lines 15-17. Note that to prevent the agents from getting stuck in a bad local optimum trap near the initial point, we add noise, $\mathcal{M}_b^t$ to the selected actions to ensure that all actions are explored. In other words, we applied Ornstein-Uhlenbeck process in order to generate noise $\mathcal{M}_b^t$ to be added to the output action of actor network to ensure that all possible actions are explored.

VI. COMPUTATIONAL COMPLEXITY ANALYSIS

The computational complexity of our proposed algorithm consists of two main parts:1) the action selection and 2) the training process.

A. COMPUTATIONAL COMPLEXITY OF ACTION SELECTION

We assume that our neural network is a fully connected neural network with fixed number of hidden layers and fixed number of neurons in each hidden layer. The computational complexity of calculating the output of such neural network for an input is equal to the sum of the sizes of input and output [67]. For our proposed algorithm, based on state and action spaces, for each BS, the sizes of the inputs of the critic and actor networks are $RK + K$ and $2RK$, respectively. K is the number of services and R is the number of RBs. Thus, the computational complexity of transmit power selection and estimation of the Q-function value for a state-action pair is $\mathcal{O}(RK)$. The estimation of the Q-function values should be done at all B BSs, thus, the computational complexity of action selection is $\mathcal{O}(BRK)$.

Algorithm 1 MA-CA2C Algorithm

1. Initialize exploration parameter; critic network parameter and actor network parameter randomly [64] for each BS b.
2. Initialize target networks parameters randomly [64] for each BS b.
3. Initialize reply buffer length D_b for each BS b.
4. for *episode from 1 to number of episodes* **do**
  5. for *cycle from 1 to number of cycles* **do**
    6. for *b from 1 to B* **do**
      7. Receive initial state s^t.
      8. With probability ϵ, select a random RB, and otherwise select an RB by policy for agent b.
      9. Determine the transmit power for the selected RB.
      10. Execute the agents actions, observe the reward and transits to new state s^{t+1}.
      11. if *all the optimization problem constraints are satisfied in state* s_b^{t+1} **then**
        12. Save transition s_b^t, a_b^t, s_b^{t+1}, r_b^t in D_b.
      end
    end
    13. if *number of transitions in* D_b *is greater than batch size* **then**
      14. Sample a random minibatch of M transitions s_b^t, a_b^t, s_b^{t+1}, r_b^t from $\mathcal{D}_b$
    end
    15. Update critic network by minimizing the loss function.
    16. Update the actor network by the sampled policy gradient.
    17. Update the target networks.
  end
end

B. COMPLEXITY OF TRAINING PROCESS

The Q-function values of the K services should be calculated and compared by the BSs before the training step. Based

on previous section, the computational complexity of this step is $\mathcal{O}(MBRK)$ where M is the size of the training batch. In addition, for a fully connected neural network in which the number of hidden layers and neurons are fixed, the back-propagation algorithm complexity is related to the product of the input size and the output size. Based on the state and action spaces, for each BS, the sizes of the inputs of the critic and actor networks are $RK + K$ and $2RK$, respectively. Moreover, for each BS, the sizes of the outputs of the critic and actor networks are 1 and 2, respectively. Thus, the back-propagation algorithm complexity is $\mathcal{O}(MBRK)$. Finally, the training process complexity is $\mathcal{O}(MBRK)$.

VII. CONVERGENCE ANALYSIS

The considered DRL algorithm, i.e., CA2C method, is an extended version of Q-learning algorithm. For Q-learning algorithm, if $\sum_{t=0}^{\infty}\alpha^t = \infty$ and $\sum_{t=0}^{\infty}(\alpha^t)^2 < \infty$ are satisfied and $|r^t(s^t, a^t)|$ is bounded, the Q-function converges to the optimal Q-function as $t \to \infty$ with probability 1 [68]. An effective approach to train neural networks is using the inverse time decaying learning rate in which using a large learning rate in the first training epochs prevents the network from getting stuck in a bad local optimum trap near the initial point [69]. Whereas, using a small learning rate in the last training epochs converges the network to a good local optimum and prevents the network from oscillation. We also analyze the convergence of our proposed algorithm through simulations in Section VIII.

VIII. SIMULATION RESULTS

In this section, the performance of our proposed frame structure with IA scheme is compared with two low complexity and overhead schemes using multi-agent and single-agent of C2AC DRL methods, defined as MA-CA2C and CA2C, respectively. It is worth noting that although there are some existing works that considered D-RBS for their framework, we are not able to compare their methods with our proposed multi-cell and multi-agent D-RBS framework, since we not only consider inter-cell interference for our formulation, but also our optimization problem is completely different. We evaluate the BSs cooperation to mitigate the impacts of inter-cell interference, as well as the impact of the URLLC packet arrival rate to the data rate of eMBB and the number of connected mMTC users [58]. In addition, our proposed IA method is also compared and evaluated. The objective of our problem is to minimize the total transmit power and the total number of allocated RBs. For simulation parameters, we assume there are 10 eMBB users, 200 mMTC users, and 5 URLLC users which are independently dispersed under an uniform distribution over a circular area with a 1000 meters radius and change their position over time according to the random walk model [1]. The traffic model for eMBB and mMTC users is considered as full buffer traffic, while for the URLLC arrival rate, it has 50 Bytes as the size of each packet and arrives at each mini-slot based on Poisson distribution [6], [58], [59]. Based on the size of URLLC packets, the BSs must allocate (puncture) 17 numbers of 1×4 resource blocks to ensure the URLLC requirements. This would increase 5.18% overhead at each mini-slot. Since we have 7 mini-slots at each decision epoch, therefore the total overhead of IA compared to RaA and OA is $7 \times 5.18\% = 36.3\%$. The other simulation parameters are summarized in Table 2.

TABLE 2. Network parameters [58], [59], [60], [61].

Parameter	Description	Value
$K^{\mathrm{e}}/K^{\mathrm{m}}/K^{\mathrm{u}}$	Number of eMBB/mMTC/URLLC users	10/200/5
f_c	Carrier frequency	2 GHz
$P_b^{\max}$	Maximum transmit power per BS [6]	40 W
$C^{\min,\mathrm{n}}$	Minimum required number of connected mMTC users	150
$R^{\min,\mathrm{e}}$	Minimum required rate for eMBB users	8 Mbps
δ	Time slot interval	0.5 ms
α	Initial learning rate	0.001
γ	Discount factor	0.99
W	Bandwidth	20 MHz
N_0	Noise power spectral	-170 dBm/Hz
κ	Path loss component	3.5
χ	Each RE time duration	0.0714 ms
ϑ	Each RE frequency size	15 kHz
–	Number of mini-slots per slot	7
F	Number of RBs	106
B	Number of BSs	3
–	Activation function for input and hidden layers	ReLU
–	Activation function for output layer	tanh
–	Optimizer	Adam
–	Number of iterations	100000
–	Number of hidden layers	3
–	Target network update frequency	1000
–	Batch size	64
–	Number of neurons in each layer	512

The simulations were conducted using Python and TensorFlow on a host PC equipped with an Intel Core i7 8th Gen CPU operating at a clock frequency of 2400 GHz. The PC had 12 GB of memory available for the simulations. Fig. 3 shows the average accumulated reward with oscillation areas (the shaded blurred colors) for our proposed model compared RaA and OA schemes in multi-agent (with -MA suffix) and single-agent (with -SA suffix) frameworks.

In terms of reward performance, Fig. 3a demonstrates that the IA approach exhibits faster convergence compared to RaA and OA methods, requiring fewer training steps (around 25000 steps for the multi-agent algorithm and 27000 steps for the single-agent algorithm). This faster convergence is attributed to IA's ability to strategically select the most suitable RBs to fulfill the objective function, facilitating more efficient progress. However, it is important to note that IA incurs an overhead of 36.3% compared to RaA and OA due to the additional processing involved in selecting the optimal RB for URLLC users. When comparing the time complexity between single-agent and multi-agent algorithms, it is evident that, in general, the single-agent algorithm requires more time

to converge for all IA, RA, and OA approaches. However, it achieves a higher reward. Specifically, when comparing IA-SA and IA-MA, the single-agent algorithm exhibits a convergence latency of approximately 2000 steps with a reward that is approximately 36% higher.

Based on the observations from Fig. 3b and Fig. 3c, it is evident that when there are 5 URLLC users with a packet arrival rate exceeding 3, the multi-agent OA (OA-MA) approach fails to provide sufficient data rate for eMBB users and connectivity for mMTC users. Similarly, the single-agent OA (OA-SA) approach struggles to meet the needs of all users when the packet arrival rate is 5. These findings highlight the lack of scalability in the OA scheme, despite some performance improvements it offers. In addition, the RaA method is less scalable compared to the IA using both single-agent and multi-agent algorithms, primarily due to higher power and RB consumption required to cater to a specific number of users across all three services. The shaded blurred regions show the oscillation of actual reward value for each method. For more clearance, we also plot the moving average of each method's reward (dashed lines for SA and solid lines for MA). It is worth noting that since our environment consists of both continuous and discrete actions, we only apply C2AC DRL method as a solution for our environment. As demonstrated, the single-agent method outperforms the multi-agent ones for all schemes. The superiority of the single-agent method over the multi-agent has the signaling overhead cost. Considering single-agent method, it can be concluded that our proposed IA-SA method achieves higher accumulated reward compared to OA-SA and RaA-SA methods with fewer training steps. This is similar for multi-agent method and it is expected due to intelligent puncturing of eMBB and mMTC RBs to URLLC users that results to consume fewer RBs and transmit power to satisfy the objective function. The lower reward for the RaA scheme compared to OA ones, apart form considering the single-agent or multi-agent methods, is because of constraint violation by RaA method. Although, the OA and RaA methods both puncture the already allocated eMBB and mMTC RBs randomly, the OA satisfies eMBB data rate and mMTC connected devices by overallocating more RBs to the eMBB and mMTC users. Fig. 4 demonstrates a sample of RB allocation for three different schemes.

As depicted, after allocation of the RBs to eMBB and mMTC users, i.e., the peach and green color REs, the URLLC packets arrived at the first, fifth, and sixth mini-slots. Then, all three methods punctures the RBs for URLLC services. Note that the IA scheme punctures the already allocated RBs of eMBB and mMTC to URLLC more properly so that the number of punctured RBs (the gray REs) is less than that of RaA and OA schemes. It is also worth mentioning that the OA and RaA methods consider the same puncturing scheme (random puncturing) for URLLC allocation. The only difference between these two methods is that the OA method allocates extra RBs to eMBB and mMTC users to compensate the punctured REs. Fig. 3(b) shows the transmit power for

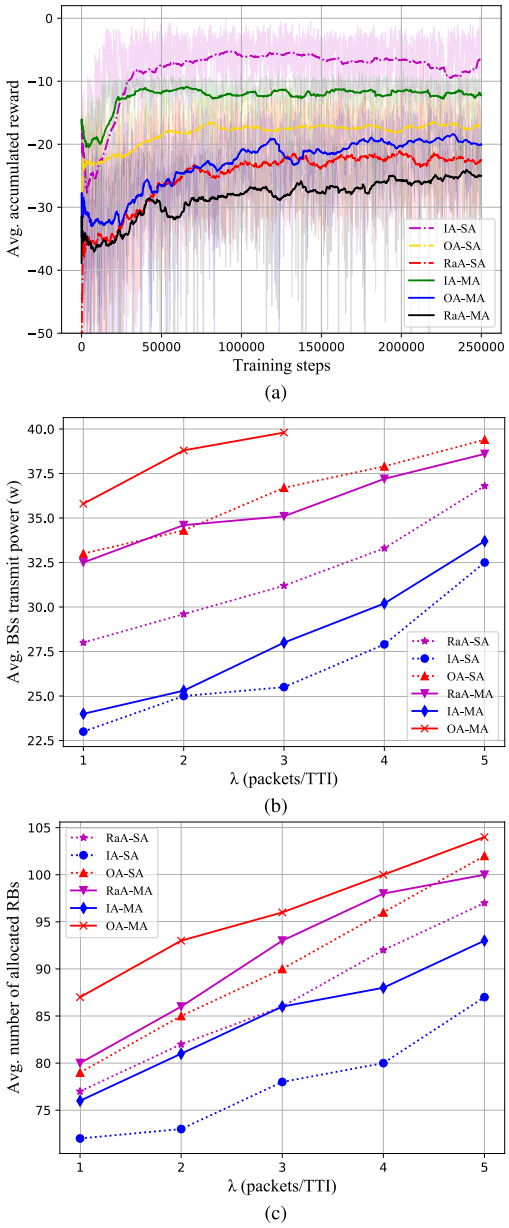


FIGURE 3. (a) The average accumulated reward for all puncturing schemes with single-agent and multi-agent methods. (b) The average BSs transmit power versus the URLLC arrival rate at each time slot. (c) The average number of allocated RBs versus eMBB required rate.

different puncturing schemes versus the URLLC arrival rate at each time slot. The best performance is for our proposed IA with single-agent DRL method. The performance gain for our proposed IA method with single-agent DRL method is 18.7% and 30.4% compared to RaA and OA methods with single-agent DRL method, respectively. In addition, the performance gain of using single-agent DRL method compared to multi-agent one for our proposed IA method is 6%. This gain for RaA and OA methods are 10% and 7.2%, respectively. Similarly, the number of allocated RB to the eMBB and mMTC for our proposed IA is the least for both single-agent and multi-agent methods.

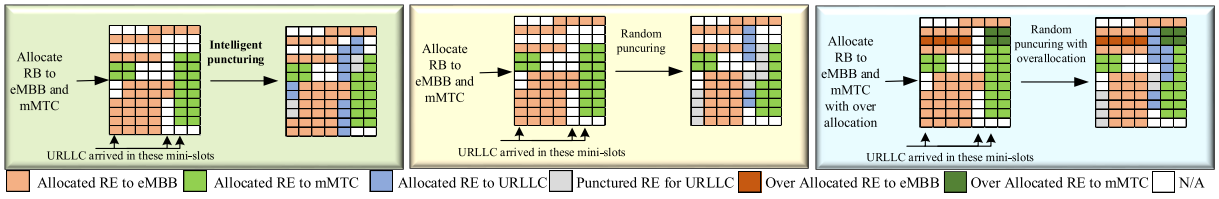


FIGURE 4. A sample of RB allocation with the three puncturing schemes: a) In the IA scheme, eMBB (Orange color), mMTC (Green color), and URLLC (black color) RBs are allocated intelligently among eMBB, mMTC, and URLLC users, respectively, where URLLC (black color) RBs are punctured intelligently. b) In RaA scheme, URLLC RBs (Gray color) are punctured randomly. C) In the OA scheme, additional RBs are allocated to eMBB (Dark orange color) and mMTC (Dark green color) users to compensate the negative effect of puncturing on data rates of eMBB and mMTC users.

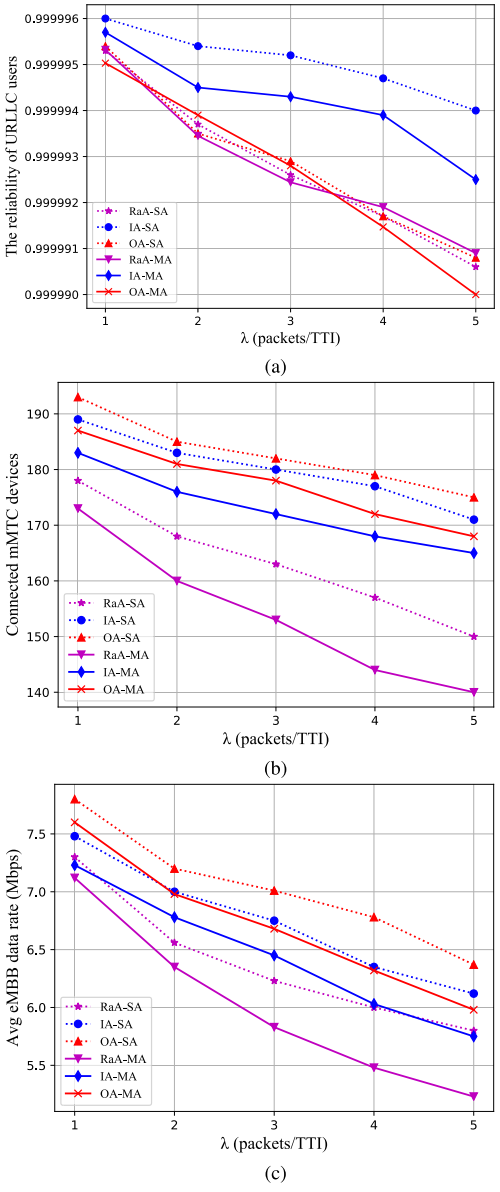


FIGURE 5. (a) The reliability versus the URLLC service packet arrival rate. (b) Total number of connected mMTC devices. (c) The average eMBB data rate.

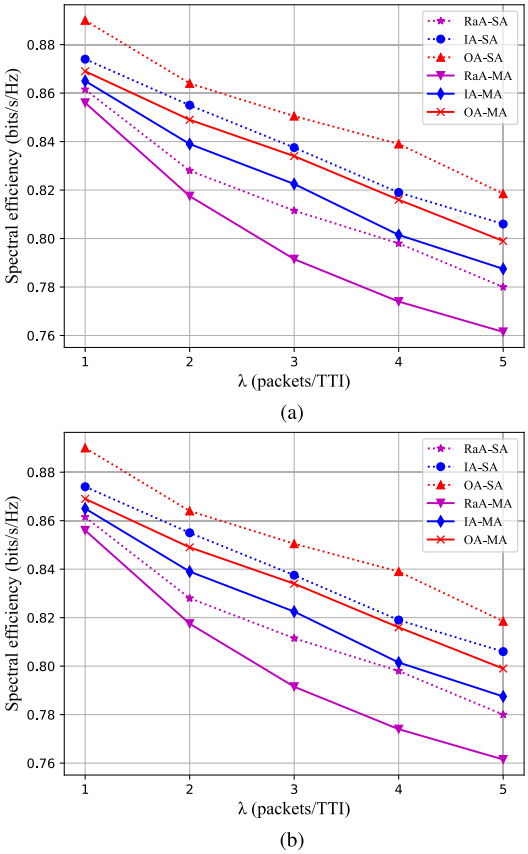


FIGURE 6. (a) The energy efficiency versus the URLLC service packet arrival rate. (b) The spectral efficiency versus the URLLC service packet arrival rate.

Fig. 5(a) depicts the reliability of the URLLC services versus the arrival of URLLC packets at each time slot. Please note that the reliability is calculated as $1 - \Phi_{lost}/\Phi_{Transmitted}$, where Φ_{lost} and $\Phi_{Transmitted}$ denote the number of lost URLLC packets and the number of transmitted URLLC packets, respectively. As shown, the increasing of the arrival packet rate at each time slot causes to decrease the reliability. The point here is all methods, apart from applying the single-agent or multi-agent method, satisfy the reliability constraint for URLLC services. The number of connected mMTC devices versus the URLLC arrived rate is demonstrated in Fig. 5(b). As expected, the OA scheme with single-agent method results in the maximum number of the connected mMTC devices compared to IA and RaA

TABLE 3. Performance and complexity of our proposed IA scheme versus OA and RaA.

Schemes	Transmit power improvement	RB usage improvement	Complexity
OA	29.2%	16.7%	36.3%
RaA	17.8%	+13%	36.3%

methods. This performance gain has come at higher transmit power and the number of allocated RBs. We will later discuss about te transmit power and allocated RBs for all methods. The average data rate by eMBB users is depicted in Fig. 5(c). Similar to Fig. 5(b), by increasing the URLLC arrival rate, the average eMBB data rate decreases. Again, the OA method achieves the highest average eMBB data rate compared to IA and RaA methods. In addition, performance and complexity of our proposed scheme are compared with both OA and RaA schemes in Table 3.

Fig. 6(a) and Fig. 6(b) show the energy efficiency and spectral efficiency (here we considered all data rate for all types of users) versus the URLLC packet arrival rate. It is concluded that the IA method with single-agent algorithm has increased the energy efficiency by 15% compared to IA with multi-agent one. This is with the cost of excessive signaling overhead in single-agent algorithm. On the other hand, our IA has increased the energy efficiency by 38% and 40% compared to RaA and OA with multi-agent algorithm, respectively. In addition, our IA approach has increased the spectral efficiency by 4% and 3.8% compared to RaA method with single-agent and multi-agent algorithms, respectively. On the other hand, OA method has increased spectral efficiency by 1.5% and 1.4% compared to IA in single-agent and multi-agent algorithms, respectively. Although the difference is not too high, this is with the cost of excessive power and resource consumption at the BSs.

IX. CONCLUSION

In this paper, we devised a novel AI-assisted dynamic RB structure abbreviated as D-RBS and introduced IA scheme for multiplexing of eMBB, mMTC, and URLLC users so that 5G networks can handle low latency traffics and large fluctuations in data rates and support heterogeneous services more efficiently. Unlike most previous works that considered dynamic RB structure in a single cell network, in this paper, we designed a new dynamic RB structure in multi-cell network with inter-cell interference which is one of main challenges in multi-cell. In addition, unlike most previous works that only consider one of latency and reliability constraints, in this paper, we consider both latency and reliability multiplexing eMBB, mMTC, and URLLC users. Furthermore, we compared this scheme with two low complexity and overhead schemes, named RaA and OA. We formulate our proposed D-RBS method with puncturing schemes as three different optimization problems in which the objectives are minimizing the network resource usage along with satisfying the eMBB and mMTC QoS requirements, while the URLLC latency and reliability constraints are satisfied. We have formulated a joint transmit power and RBs assignment problem to minimize the long-term energy consumption and the number of allocated RBs. We developed a DRL-based algorithm to solve our optimization problem with both discrete and continuous actions. Simulation results have verified our algorithms convergence and showed that our proposed scheme, i.e., D-RBS, could achieve much better performance than the traditional scheme, S-RBS. Since our optimization problems are nonlinear non-convex with multiple discrete and continuous variables, we used a DRL method to optimize the transmit power allocation and RBs assignment in single-agent and multi-agent scenarios and showed that the proposed framework scales well with a large number of RBs, BSs, and users. Then, we analyzed the performance gain of these three schemes which comprehensively indicated the trade-off between network resource usage and user's satisfaction. We further investigated our proposed schemes from the convergence and computational complexity perspectives. Finally, we studied the performance of the proposed schemes and compared it with S-RBS baseline approach using simulations for different network parameters. We showed that our proposed IA scheme achieves a performance gain of 30% and 60% compared to the RaA and OA schemes, respectively. However, IA had 36.3% complexity for action selection compared to both RaA and OA.

REFERENCES

[1] *Study on Scenarios and Requirements for Next Generation Access Technologies, Technical Specification Group Radio Access Network*, 3GPP, document TS 38.913, Oct. 2016.
[2] (2017). *Minimum Requirements Related to Technical Performance for IMT-2020 Radio Interface(s)*. [Online]. Available: https://www.itu.int/dmspub/itu-r/opb/rep/R-REP-M.2410-2017-PDF-E.pdf
[3] *Study on Physical Layer Enhancements for NR Ultra-Reliable and Low Latency Case (URLLC)*, 3GPP, document TR 38.824, 2017. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3498
[4] (Dec. 1, 2017). *Final Report of 3GPP TSG RAN WG1 Meeting 91 version 1.0.0*. Reno, NV, USA. [Online]. Available: https://www.3gpp.org/ftp/tsgran/WG1RL1/TSGR191/Report/
[5] M. Alsenwi, N. H. Tran, M. Bennis, S. R. Pandey, A. K. Bairagi, and C. S. Hong, "Intelligent resource slicing for eMBB and URLLC coexistence in 5G and beyond: A deep reinforcement learning based approach," *IEEE Trans. Wireless Commun.*, vol. 20, no. 7, pp. 4585–4600, Jul. 2021.
[6] Y. Huang, S. Li, C. Li, Y. T. Hou, and W. Lou, "A deep-reinforcement-learning-based approach to dynamic eMBB/URLLC multiplexing in 5G NR," *IEEE Internet Things J.*, vol. 7, no. 7, pp. 6439–6456, Jul. 2020.
[7] L. You, Q. Liao, N. Pappas, and D. Yuan, "Resource optimization with flexible numerology and frame structure for heterogeneous services," *IEEE Commun. Lett.*, vol. 22, no. 12, pp. 2579–2582, Dec. 2018.
[8] W. Sui, X. Chen, S. Zhang, Z. Jiang, and S. Xu, "Energy-efficient resource allocation with flexible frame structure for hybrid eMBB and URLLC services," *IEEE Trans. Green Commun. Netw.*, vol. 5, no. 1, pp. 72–83, Mar. 2021.
[9] G. Wang, G. Feng, W. Tan, S. Qin, R. Wen, and S. Sun, "Resource allocation for network slices in 5G with network resource pricing," in *Proc. IEEE Global Commun. Conf.*, Dec. 2017, pp. 1–6.
[10] R. Su, D. Zhang, R. Venkatesan, Z. Gong, C. Li, F. Ding, F. Jiang, and Z. Zhu, "Resource allocation for network slicing in 5G telecommunication networks: A survey of principles and models," *IEEE Netw.*, vol. 33, no. 6, pp. 172–179, Nov. 2019.

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

[12] K. I. Pedersen, G. Berardinelli, F. Frederiksen, P. Mogensen, and A. Szufarska, "A flexible 5G frame structure design for frequency-division duplex cases," *IEEE Commun. Mag.*, vol. 54, no. 3, pp. 53–59, Mar. 2016.

[13] S.-Y. Lien, S.-C. Hung, D.-J. Deng, and Y. J. Wang, "Efficient ultra-reliable and low latency communications and massive machine-type communications in 5G new radio," in *Proc. IEEE Global Commun. Conf.*, Singapore, Dec. 2017, pp. 1–7.

[14] C. Sun, C. She, and C. Yang, "Energy-efficient resource allocation for ultra-reliable and low-latency communications," in *Proc. IEEE Global Commun. Conf.*, Singapore, Dec. 2017, pp. 1–6.

[15] M. Shafi, A. F. Molisch, P. J. Smith, T. Haustein, P. Zhu, P. De Silva, F. Tufvesson, A. Benjebbour, and G. Wunder, "5G: A tutorial overview of standards, trials, challenges, deployment, and practice," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 6, pp. 1201–1221, Jun. 2017.

[16] Z. Dawy, W. Saad, A. Ghosh, J. G. Andrews, and E. Yaacoub, "Toward massive machine type cellular communications," *IEEE Wireless Commun.*, vol. 24, no. 1, pp. 120–128, Feb. 2017.

[17] H. Zhang, N. Liu, X. Chu, K. Long, A.-H. Aghvami, and V. C. M. Leung, "Network slicing based 5G and future mobile networks: Mobility, resource management, and challenges," *IEEE Commun. Mag.*, vol. 55, no. 8, pp. 138–145, Aug. 2017.

[18] J. Khan and L. Jacob, "Resource allocation for CoMP enabled URLLC in 5G C-RAN architecture," *IEEE Syst. J.*, vol. 15, no. 4, pp. 4864–4875, Dec. 2021.

[19] N. H. Mahmood, M. Lauridsen, G. Berardinelli, D. Catania, and P. Mogensen, "Radio resource management techniques for eMBB and mMTC services in 5G dense small cell scenarios," in *Proc. IEEE 84th Veh. Technol. Conf. (VTC-Fall)*, Montreal, QC, Canada, Sep. 2016, pp. 1–5.

[20] L. Tian, C. Yan, W. Li, Z. Yuan, W. Cao, and Y. Yuan, "On uplink non-orthogonal multiple access for 5G: Opportunities and challenges," *China Commun.*, vol. 14, no. 12, pp. 142–152, Dec. 2017.

[21] C. She, C. Yang, and T. Q. S. Quek, "Cross-layer optimization for ultra-reliable and low-latency radio access networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 127–141, Jan. 2018.

[22] A. Aijaz, "Towards 5G-enabled tactile internet: Radio resource allocation for haptic communications," in *Proc. IEEE Wireless Commun. Netw. Conf. Workshops (WCNCW)*, Doha, Qatar, Apr. 2016, pp. 145–150.

[23] C. She, C. Yang, and T. Q. S. Quek, "Radio resource management for ultra-reliable and low-latency communications," *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 72–78, Jun. 2017.

[24] C. She, C. Yang, and T. Q. S. Quek, "Joint uplink and downlink resource configuration for ultra-reliable and low-latency communications," *IEEE Trans. Commun.*, vol. 66, no. 5, pp. 2266–2280, May 2018.

[25] P. Popovski, K. F. Trillingsgaard, O. Simeone, and G. Durisi, "5G wireless network slicing for eMBB, URLLC, and mMTC: A communication-theoretic view," *IEEE Commun. Lett.*, vol. 24, no. 3, pp. 658–661, Dec. 2019.

[26] B. Matthiesen, O. Aydin, and E. A. Jorswieck, "Throughput and energy-efficient network slicing," in *Proc. 22nd Int. ITG Workshop Smart Antennas*, Bochum, Germany, Mar. 2018, pp. 1–6.

[27] Z. Wu, F. Zhao, and X. Liu, "Signal space diversity aided dynamic multiplexing for eMBB and URLLC traffics," in *Proc. 3rd IEEE Int. Conf. Comput. Commun. (ICCC)*, Chengdu, China, Dec. 2017, pp. 1396–1400.

[28] M. Iwabuchi, A. Benjebbour, Y. Kishiyama, G. Ren, C. Tang, T. Tian, L. Gu, T. Takada, and T. Kashima, "5G field experimental trials on URLLC using new frame structure," in *Proc. IEEE Globecom Workshops*, Singapore, Dec. 2017, pp. 1–6.

[29] Q. Liao, P. Baracca, D. Lopez-Perez, and L. G. Giordano, "Resource scheduling for mixed traffic types with scalable TTI in dynamic TDD systems," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Washington, DC, USA, Dec. 2016, pp. 1–7.

[30] E. Fountoulakis, N. Pappas, Q. Liao, V. Suryaprakash, and D. Yuan, "An examination of the benefits of scalable TTI for heterogeneous traffic management in 5G networks," in *Proc. 15th Int. Symp. Modeling Optim. Mobile, Ad Hoc, Wireless Netw. (WiOpt)*, Paris, France, May 2017, pp. 1–6.

[31] X. Jiang, K. Liang, X. Chu, C. Li, and G. K. Karagiannidis, "Multiplexing eMBB and URLLC in wireless powered communication networks: A deep reinforcement learning-based approach," *IEEE Wireless Commun. Lett.*, vol. 12, no. 10, pp. 1716–1720, Jun. 2023.

[32] H. Peng, T. Kallehauge, M. Tao, and P. Popovski, "Power and rate adaptation for URLLC with statistical channel knowledge and HARQ," *IEEE Wireless Commun. Lett.*, early access, Aug. 30, 2023, doi: 10.1109/LWC.2023.3310205.

[33] H. Han, X. Jiang, W. Lu, W. Zhai, Y. Li, N. Kumar, and M. Guizani, "A multi-agent reinforcement learning approach for massive access in NOMA-URLLC networks," *IEEE Trans. Veh. Technol.*, early access, Jul. 6, 2023, doi: 10.1109/TVT.2023.3292423.

[34] J. Deng, O. Tirkkonen, R. Freij-Hollanti, T. Chen, and N. Nikaein, "Resource allocation and interference management for opportunistic relaying in integrated mmWave/sub-6 GHz 5G networks," *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 94–101, Jun. 2017.

[35] L. Huo and D. Jiang, "Stackelberg game-based energy-efficient resource allocation for 5G cellular networks," *Telecommun. Syst.*, vol. 72, no. 3, pp. 377–388, Nov. 2019.

[36] Q. Chen, J. Wu, J. Wang, and H. Jiang, "Coexistence of URLLC and eMBB services in MIMO-NOMA systems," *IEEE Trans. Veh. Technol.*, vol. 72, no. 1, pp. 839–851, Jan. 2023.

[37] H. Peng, L.-C. Wang, and Z. Jian, "Data-driven spectrum partition for multiplexing URLLC and eMBB," *IEEE Trans. Cognit. Commun. Netw.*, vol. 9, no. 2, pp. 386–397, Apr. 2023.

[38] D. E. Pérez, O. L. Alcaraz López, and H. Alves, "Robust downlink multi-antenna beamforming with heterogenous CSI: Enabling eMBB and URLLC coexistence," *IEEE Trans. Wireless Commun.*, vol. 22, no. 2, pp. 4146–4157, Nov. 2022.

[39] G. S. Kesava and N. B. Mehta, "Multi-connectivity for URLLC and coexistence with eMBB in time-varying and frequency-selective fading channels," *IEEE Trans. Wireless Commun.*, vol. 22, no. 6, pp. 3599–3611, Nov. 2022.

[40] H. Ye, G. Y. Li, and B. F. Juang, "Deep reinforcement learning based resource allocation for V2V communications," *IEEE Trans. Veh. Technol.*, vol. 68, no. 4, pp. 3163–3173, Apr. 2019.

[41] Y. Wei, F. R. Yu, M. Song, and Z. Han, "User scheduling and resource allocation in HetNets with hybrid energy supply: An actor-critic reinforcement learning approach," *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 680–692, Jan. 2018.

[42] A. T. Nassar and Y. Yilmaz, "Reinforcement-learning-based resource allocation in fog radio access networks for various IoT environments," 2018, Art. no. arXiv:1806.04582.

[43] Y. Yu, T. Wang, and S. C. Liew, "Deep-reinforcement learning multiple access for heterogeneous wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1277–1290, Jun. 2019.

[44] S. Wang, H. Liu, P. H. Gomes, and B. Krishnamachari, "Deep reinforcement learning for dynamic multichannel access in wireless networks," *IEEE Trans. Cognit. Commun. Netw.*, vol. 4, no. 2, pp. 257–265, Jun. 2018.

[45] O. Naparstek and K. Cohen, "Deep multi-user reinforcement learning for distributed dynamic spectrum access," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 310–323, Jan. 2019.

[46] S. Liu, X. Hu, and W. Wang, "Deep reinforcement learning based dynamic channel allocation algorithm in multibeam satellite systems," *IEEE Access*, vol. 6, pp. 15733–15742, 2018.

[47] K. A. Yau, P. Komisarczuk, and D. T. Paul, "Enhancing network performance in distributed cognitive radio networks using single-agent and multi-agent reinforcement learning," in *Proc. IEEE Local Comput. Netw. Conf.*, Oct. 2010, pp. 152–159.

[48] Z. Yang, K. Merrick, L. Jin, and H. A. Abbass, "Hierarchical deep reinforcement learning for continuous action control," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 11, pp. 5174–5184, Nov. 2018.

[49] K. Salah, K. Elbadawi, and R. Boutaba, "An analytical model for estimating cloud resources of elastic services," *J. Netw. Syst. Manage.*, vol. 24, no. 2, pp. 285–308, Apr. 2016.

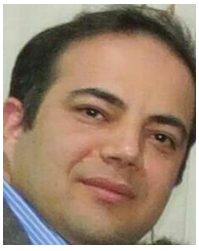
[50] S. Zheng and H. Liu, "Improved multi-agent deep deterministic policy gradient for path planning-based crowd simulation," *IEEE Access*, vol. 7, pp. 147755–147770, 2019.

[51] L. Panait and S. Luke, "Cooperative multi-agent learning: The state of the art," *Auto. Agents Multi-Agent Syst.*, vol. 11, no. 3, pp. 387–434, Nov. 2005.

[52] M. C. Lucas-Estañ and J. Gozalvez, "Load balancing for reliable self-organizing industrial IoT networks," *IEEE Trans. Ind. Informat.*, vol. 15, no. 9, pp. 5052–5063, Sep. 2019.

[53] S. K. C. Lo, "A collaborative multi-agent message transmission mechanism in intelligent transportation system—A smart freeway example," *Inf. Sci.*, vol. 184, no. 1, pp. 246–265, Feb. 2012.

[54] *5G; Study on New Radio Access Technology*, 3GPP, document TR 138 912, 2017.
[55] G. Durisi, T. Koch, and P. Popovski, "Toward massive, ultrareliable, and low-latency wireless communication with short packets," *Proc. IEEE*, vol. 104, no. 9, pp. 1711–1726, Sep. 2016.
[56] W. Yang, G. Durisi, T. Koch, and Y. Polyanskiy, "Quasi-static multiple-antenna fading channels at finite blocklength," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 4232–4265, Jul. 2014.
[57] C. She, Z. Chen, C. Yang, T. Q. S. Quek, Y. Li, and B. Vucetic, "Improving network availability of ultra-reliable and low-latency communications with multi-connectivity," *IEEE Trans. Commun.*, vol. 66, no. 11, pp. 5482–5496, Nov. 2018.
[58] X. Zhang, J. Wang, and H. V. Poor, "Statistical delay and error-rate bounded QoS provisioning for mURLLC over 6G CF M-MIMO mobile networks in the finite blocklength regime," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 3, pp. 652–667, Mar. 2021.
[59] X. Zhang, J. Wang, and H. V. Poor, "Optimal resource allocations for statistical QoS provisioning to support mURLLC over FBC-EH-based 6G THz wireless nano-networks," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 6, pp. 1544–1560, Jun. 2021.
[60] E. Markoval, D. Moltchanov, R. Pirmagomedov, D. Ivanova, Y. Koucheryavy, and K. Samouylov, "Priority-based coexistence of eMBB and URLLC traffic in industrial 5G NR deployments," in *Proc. 12th Int. Congr. Ultra Mod. Telecommun. Control Syst. Workshops (ICUMT)*, Brno, Czech Republic, Oct. 2020, pp. 1–6.
[61] A. Anand and G. de Veciana, "Resource allocation and HARQ optimization for URLLC traffic in 5G wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 11, pp. 2411–2421, Nov. 2018.
[62] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *J. Artif. Intell. Res.*, vol. 4, no. 1, pp. 237–285, Jan. 1996.
[63] T. M. Moerland, J. Broekens, A. Plaat, and C. M. Jonker, "Model-based reinforcement learning: A survey," *Found. Trends® Mach. Learn.*, vol. 16, no. 1, pp. 1–118, 2023.
[64] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. London, U.K.: MIT Press, 2018.
[65] J. Hu, H. Zhang, L. Song, R. Schober, and H. V. Poor, "Cooperative Internet of UAVs: Distributed trajectory design by multi-agent deep reinforcement learning," *IEEE Trans. Commun.*, vol. 68, no. 11, pp. 6807–6821, Nov. 2020.
[66] R. Lowe, Y. I. Wu, A. Tamar, J. Harb, O. P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6379–6390.
[67] M. Sipper, "A serial complexity measure of neural networks," in *Proc. IEEE Int. Conf. Neural Netw.*, San Francisco, CA, USA, Aug. 1993, pp. 962–966.
[68] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Mach. Learn.*, vol. 8, nos. 3–4, pp. 279–292, 1992.
[69] K. You, M. Long, J. Wang, and M. I. Jordan, "How does learning rate decay help modern neural networks?" 2019, *arXiv:1908.01878*.



MOHAMMAD REZA ABEDI (Student Member, IEEE) received the M.Sc. degree in electrical engineering from Amirkabir University, Tehran, Iran. He is currently pursuing the Ph.D. degree with Tarbiat Modares University, Tehran. He is also a Research Assistant with Tarbiat Modares University. He has been involved in a number of large scale network design and consulting projects in the telecom industry, as a Principle Investigator or a Consultant. His research interests include multiple access techniques, energy harvesting and wireless power transfer, cooperative and adaptive wireless communications, wireless edge caching, mobile edge computing, multibitrate video transcoding, software defined networking, wireless network virtualization, and optimization theory. He was a member of Technical Program Committees for the IEEE Conferences. He was a reviewer of several IEEE journals, such as IEEE TRANSACTIONS ON SIGNAL PROCESSING and IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS.

MOHAMMAD REZA JAVAN (Senior Member, IEEE) received the B.Sc. degree in electrical engineering from Shahid Beheshti University, Tehran, Iran, in 2003, the M.Sc. degree in electrical engineering from the Sharif University of Technology, Tehran, in 2006, and the Ph.D. degree in electrical engineering from Tarbiat Modares University, Tehran, in 2013. He is currently a Faculty Member with the Department of Electrical Engineering, Shahrood University of Technology, Shahrud, Iran. His research interests include design and analysis of wireless communication networks with emphasis on the application of optimization theory and machine learning methods.

MOHSEN POURGHASEMIAN received the B.Sc. and M.Sc. degrees in electrical engineering, in 2014 and 2016, respectively. He is currently pursuing the Ph.D. degree with the Institute for Communication Technologies and Embedded Systems, RWTH Aachen University. From August 2017 to October 2018, he was with the ICT Research Institute, Tehran, Iran, as a Researcher, where he participated in the IoT projects on 5G networks. From December 2019 to April 2022, he was a Research Assistant with Tarbiat Modares University, in the field of wireless network and machine learning. His research interests include wireless networks, resource management, machine learning, and software defined radios.

NADER MOKARI (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from Tarbiat Modares University, Tehran, Iran, in 2014. He joined the Department of Electrical and Computer Engineering, Tarbiat Modares University, as an Assistant Professor, in October 2015, where he is currently an Associated Professor. He was involved in a number of large scale network design and consulting projects in the telecom industry. His research interest includes cover many aspects of wireless technologies with a special emphasis on wireless networks. In recent years, his research has been funded by Iranian Mobile Telecommunication Companies and the Iranian National Science Foundation (INSF). His thesis received the IEEE Outstanding Ph.D. Thesis Award. He has been elected as an IEEE Exemplary Reviewer by the IEEE Communications Society, in 2016. He received the Best Paper Award by ITU K-2020. He is on the Editorial Board of IEEE TRANSACTIONS ON COMMUNICATIONS.

EDUARD A. JORSWIECK (Fellow, IEEE) was the Head of the Chair of Communications Theory and a Full Professor with TU Dresden, Germany, from 2008 to 2019. He is currently the Managing Director of the Institute of Communications Technology, the Head of the Chair for Communications Systems, and a Full Professor with Technische Universität Braunschweig, Brunswick, Germany. His research interest includes broad area of communications. He has published some 150 journal papers, 15 book chapters, three monographs, and 300 conference papers on these topics. In 2006, he received the IEEE Signal Processing Society Best Paper Award. Since 2017, he has been serving as the Editor-in-Chief for the *EURASIP Journal on Wireless Communications and Networking*. He has served on the Editorial Boards for IEEE TRANSACTIONS ON SIGNAL PROCESSING, IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE SIGNAL PROCESSING LETTERS, and IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY.