IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 13 November 2023, accepted 3 December 2023, date of publication 14 December 2023,
date of current version 3 January 2024.

Digital Object Identifier 10.1109/ACCESS.2023.3343250

SURVEY

Resource Allocation for Co-Existence of eMBB and URLLC Services in 6G Wireless Networks: A Survey

BHAGAWAT ADHIKARI, (Member, IEEE), MUHAMMAD JASEEMUDDIN, (Member, IEEE), AND ALAGAN ANPALAGAN, (Senior Member, IEEE)
Department of Electrical, Computer, and Biomedical Engineering, Toronto Metropolitan University, Toronto, ON M5B 2K3, Canada

Corresponding author: Bhagawat Adhikari (b3adhika@torontomu.ca)

The work of Muhammad Jaseemuddin and Alagan Anpalagan was supported in part by NSERC Discovery Grants.

ABSTRACT Next generation of wireless networks are characterized by two main features named Enhanced Mobile Broadband (eMBB) and Ultra Reliable Low Latency Communications (URLLC). These two services can be accommodated in the same wireless infrastructure so that wide range of users, demanding either massive throughput or extremely low latency and high reliability requirements, are directly benefited for providing various mission critical services. Co-existence of eMBB and URLLC services, however, demand highly efficient and less complex resource allocation schemes. In this paper, various resource allocation techniques are studied for the co-existence of eMBB and URLLC traffic to meet the heterogeneous specifications of each class of users. A detailed study on existing resource allocation schemes for simultaneous transmission of eMBB and URLLC services based on network slicing, flexible Transmit Time Interval (TTI), scheduling and distributed and federated learning are provided. Moreover, Machine Learning (ML) aided and Reconfigurable Intelligent Surface (RIS) and Unmanned Aerial Vehicle (UAV) assisted resource allocation techniques are also studied in detail. Additionally, this paper identifies some challenges for eMBB and URLLC service accommodations in the same wireless architecture, and proposes their possible solution approaches.

INDEX TERMS eMBB and URLLC, multiplexing, resource allocation, slicing, transmit time interval, machine learning.

I. INTRODUCTION

Future generation of communication networks, especially 6G, is expected to support a wide range of new applications and services. Rapid technological advancement emerges in contemporary wireless network industries. As a result, wide range of critical services, applications and requirements are to be addressed. These emerging applications demand enhanced and stricter requirements than those supported by previous generation networks [1]. For example, Augmented and Virtual Reality (AR and VR), Extended Reality (XR), autonomous vehicles, massive Internet of Things (IoT) devices connectivity, Artificial Intelligence (AI) based services, smart cities, smart farming and many other applications in rapid proliferation of time critical services demand specific requirements of milliseconds latency and Terabits per seconds (Tbps) data rates. For handling diverse applications and services related to latency, coverage, reliability and throughput, three service categories namely Enhanced Mobile Broadband (eMBB), Ultra Reliable Low Latency Communications (URLLC) and Massive Machine Type Communications (mMTC) need to be designed and implemented. A typical 6G heterogeneous service scenarios and use cases has been presented in Figure 1.

eMBB traffic possesses large payload and it is activated by a device with an unchanged pattern during the extension of the time period. eMBB service in 6G is an extended version of it's 5G counterpart. One of the important characteristics of eMBB is that only one eMBB device is being scheduled and served at a time so that no two eMBB devices are allowed

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

TABLE 1. Important Acronyms and their explanations.

Acronym	Explanation
AMC	Adaptive Modulation and Coding (AMC)
AR	Augmented Reality
BCD	Block Co-ordinate Descent
DRL	Deep Reinforcement Learning
DDPG	Deep Deterministic Policy Gradient
DRAPS	Dynamic Resource Allocation and Puncturing Strategy
DPP	Lyapunov Drift-plus-penalty
DL	Deep Learning
DNN	Deep Neural Networks
DRRA	Decomposition and Relaxation-based Resource Allocation
EC	Effective Capacity
eNSBPS	Enhanced Null Space Based Preemptive Scheduler
FBC	Finite Block Coding
FDD	Frequency Division Duplex
HMA	Hybrid Multiple Access (HMA)
HARQ	Hybrid Automatic Repeat Request
IoT	Internet of Things
KPI	Key Performance Indicator
MODI	Modified Distribution
MINLP	Mixed Integer Nonlinear Programming
MCC	Minimum Cell Cost (MCC)
MEAR	Minimum Expected Achieved Rate
MDP	Markov Decision Process
NLOS	Non-Line of Sight
NOMA	Non-Orthogonal Multiple Access
NFV	Network Functions Virtualization
OMA	Orthogonal Multiple Access
OFDMA	Orthogonal Frequency Division Multiple Access
PSUM	Penalty-based Successive Upper Bound Minimization
PLR	Packet Loss Rate
QoS	Quality of Service
RB	Resource Block
RAN	Radio Access Network
RF	Random Forest
RIS	Reconfigurable Intelligent Surface
RSMA	Rate Splitting Multiple Access
SIC	Successive Interference Cancellation
SW	Sliding Window
SDN	Software Defined Network
SCA	Successive Convex Approximation
SE	Spectral Efficiency
TIN	Treating Interference as Noise (TIN)
TM	Transportation Model
TTI	Transmit Time Interval
TDMA	Time Division Multiple Access
UAV	Unmanned Aerial Vehicle
VR	Virtual Reality

to access the same network resources simultaneously. eMBB service cares more about the data rate than the reliability and latency. Consequently, it has higher packet error rate. On the other hand, mMTC service is expected to serve large number of IoT devices with typically low uplink data rate. Among various IoT devices targeted to be served, only a few of them will be connected to the base station at a time. mMTC service aims to provide the maximum arrival rate that a given radio resource supports definite cluster of given IoT devices. However, packet error rate in an individual mMTC transmission is higher than that in eMBB service. URLLC transmission aims for the mission critical services that require high reliability and low latency. This service transmission is intermittent and being served within the transmission of the eMBB and mMTC services. URLLC cares less about the data rate but more about the reliability. As a result, the packet error rate with such low latency and high reliable URLLC service is typically very low of about 10^{-5} or lower.

6G wireless communication specifies eMBB and URLLC as two critical services that demand higher data rates and milliseconds latency with 99.999% reliability, respectively [2]. To provide these two services in a single wireless network, simultaneous co-existence models have been widely studied and explored in the literature [3]. eMBB and URLLC signal transmission demand specific requirements: eMBB transmission requires long data packets with higher payload that generally follow the Shannon capacity formulation whereas URLLC transmission is possible with the short data packets transmission. Consequently, URLLC aims low end-to-end latency and eMBB targets in higher data rate transmission with large bandwidth [4]. As two services are essential for fulfilling the different data rate, latency, reliability and QoS requirement of different users, these services need to be transmitted simultaneously in the same network. Consequently, the scientists and the researchers have provided various simultaneous transmission schemes in the literature. Some of the important eMBB-URLLC co-existence techniques are network slicing, puncturing and superposition. In these techniques, signals are transmitted in slots and mini slots. eMBB signals are transmitted in the slots and URLLC signals are transmitted in the mini-slots within the slots to satisfy QoS requirements of different services. In the puncturing scheme, eMBB signal is completely replaced by the URLLC signal whereas in the superposition scheme, both eMBB and uRLLC services share the resources in the mini slots and both signals are transmitted within the same mini slots [5], [6].

6G is expected to support more critical and challenging scenarios than in 5G networks. These scenarios include the use of massive number of URLLC devices. When these large number of devices will be operated under the massive URLLC (mURLLC) use case scenario, then the resource allocations to meet the diverse requirements of each use cases will be more challenging [7]. As the number of URLLC devices is expected to increase massively in few years, resource allocation needs to be optimized to lower the impact on the QoS of eMBB and mMTC users in the same network while meeting the latency and reliability requirements of massive URLLC users. Many applications of 6G such as supporting tremendous number of connected devices with rate-hungry applications such as extended reality, autonomous vehicles and wireless brain-computer interactions belong to the service-class called Mobile Broadband Reliable Low Latency Communications (MBRLLC) [8]. This class of service allows 6G networks to provide massive data rate with extremely high reliability and low latency communications, and fulfills the rate-reliability-latency constraints with consideration of energy efficiency and limitation of resources.

There are many research going on in academia that have identified various issues in 6G and have proposed

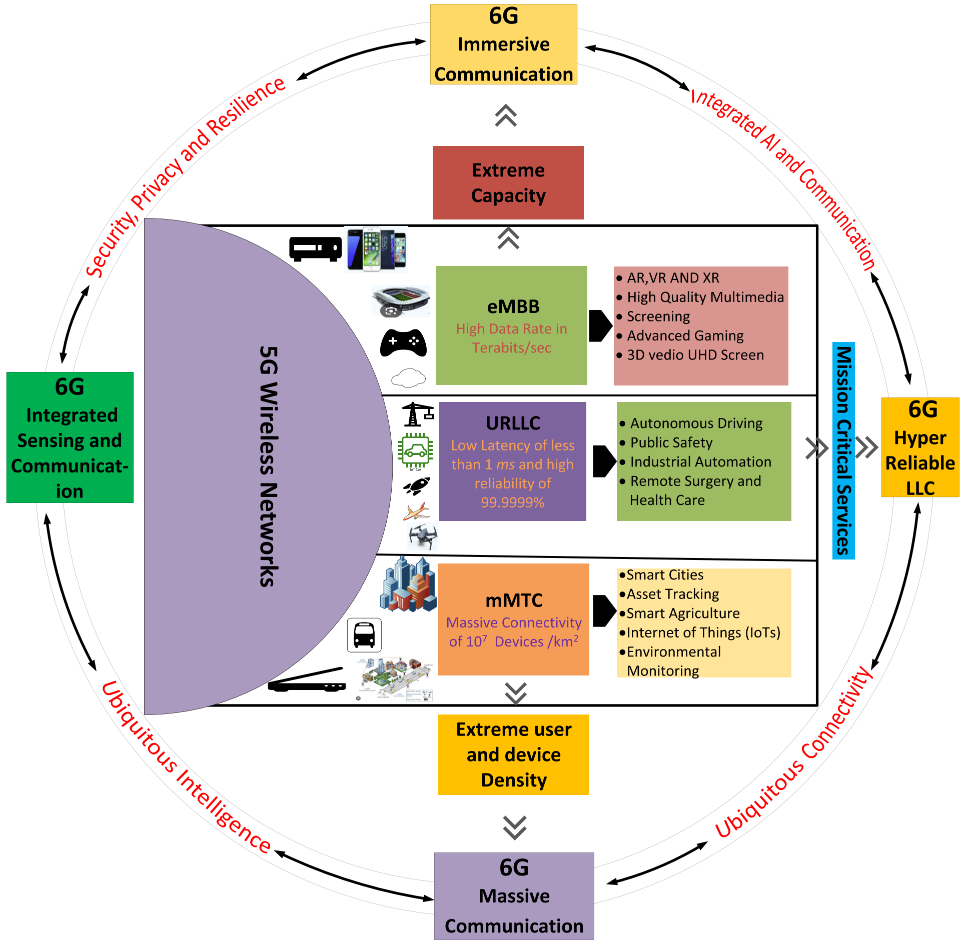

6G Immersive Communication
Integrated AI and Communication
6G Hyper Reliable LLC
Ubiquitous Connectivity
6G Massive Communication
Ubiquitous Intelligence
6G Integrated Sensing and Communicat-ion
Security, Privacy and Resilience
Extreme Capacity
5G Wireless Networks
eMBB
High Data Rate in Terabits/sec
• AR,VR AND XR
• High Quality Multimedia
• Screening
• Advanced Gaming
• 3D vedio UHD Screen
URLLC
Low Latency of less than 1 ms and high reliability of 99.9999%
• Autonomous Driving
• Public Safety
• Industrial Automation
• Remote Surgery and Health Care
mMTC
Massive Connectivity of 10^7 Devices /km^2
•Smart Cities
•Asset Tracking
•Smart Agriculture
•Internet of Things (IoTs)
•Environmental Monitoring
Mission Critical Services
Extreme user and device Density

FIGURE 1. 6G heterogeneous service scenarios and use cases.

their possible solutions. Some of the important research areas in 6G are decentralization networks, integration of satellites, terrestrial and airborne networks, AI use cases, QOPE metrics, privacy and security, 3CLS (communication, computation, control, localization and sensing), integration of heterogeneous multiple frequency bands are few to mention [8]. Resource allocation is one of the major issues in the wireless networks designed for above stated use cases due to the specific QoS requirements and heterogeneity of diverse service scenarios. As 6G demands higher data transmission rate with the latency of less than 1 ms and reliability of more than 99.999% in more complex, diverse and integrated wireless network scenarios with millions of devices connectivity per square km, resource allocation for eMBB, URLLC and mMTC service cases in 6G will be more challenging than in 5G. Additionally, target applications for 6G require the combination of various heterogeneous services, e.g. AR/VR needs eMBB and URLLC, autonomous vehicles need URLLC and mMTC etc. Moreover, resource allocation in above scenarios is more critical when the users with the various service requirements (latency, reliability and data rates) are present in the same network infrastructure. For this reason, various resource allocation techniques have been proposed, studied and analyzed for the co-existence of various service classes in 6G.

Various works have been accomplished regarding resource allocations for the co-existence of different service classes in B5G and 6G networks. Some of the resource allocation schemes such as superposition, puncturing and network slicing have been proposed in the literature. For example, in [9], the authors have studied the superposition scheme, and have outlined the benefits of using superposition for providing the uplink communication among three different service classes namely eMBB, mMTC, and URLLC. On the other hand, both downlink and uplink communications are studied in [10] where non-orthogonal co-existence of URLLC and eMBB services are analyzed. In this work, URLLC and eMBB traffic are processed at the network edge and the cloud radio access respectively so that eMBB spectral efficiency has been improved by guaranteeing the URLLC latency requirements. Another study in [11] provides a MIMO-NOMA aided URLLC traffic for downlink communication. Both heterogeneous orthogonal and non-orthogonal multiple excess network slicing techniques are used to develop a max-matching diversity algorithm to allocate the eMBB traffic in [12]. We have collected information on eMBB and URLLC services from various research papers and presented as a survey in this paper. Here, we are providing recent research on resource allocation techniques for the co-existence of eMBB and uRLLC services in 6G networks.

A. EXISTING RELATED SURVEY PAPERS

eMBB and URLLC wireless services have been studied in some survey papers in the literature. Most of these survey works are based on URLLC services. The authors in [13] have surveyed the packet scheduling algorithms in URLLC for 5G and B5G networks. A review of good number of state-of-art techniques focusing on centralized, decentralized and scheduling schemes for URLLC have been provided. Another work in [14] discusses 5G NR system with the architecture and emerging technologies in B5G networks. Most of the discussion in [14] is about the challenges concerning the 5G NR implementation where eMBB, URLLC and mMTC are discussed as important use cases. A comprehensive review on Industrial Internet of Things (IIoT) wireless networks based on the application of B5G has been provided in [15]. This work enlightens the trade-off between a number of good applications and key-enabling technologies for eMBB and URLLC. Some other surveys in [16] and [17] provide in-depth insights on applications, challenges, implemented techniques with the required latency, reliability and throughput in various use cases in IIoT based URLLC systems. The aforementioned survey works focus on either URLLC system or the separate existence of eMBB and URLLC. However, review on eMBB and URLLC co-existence in B5G and 6G is still missing in the literature. Different to previous survey works, in this paper, we will provide a detailed study on different resource allocation schemes focusing on the QoS requirements for eMBB and URLLC along with their implementation challenges. An overview of existing surveys on eMBB and URLLC co-existence has been presented in Table 2 that clearly shows the differences of this survey with the existing survey works.

B. MOTIVATION AND OBJECTIVES

Aforementioned survey papers have reviewed the existing works about eMBB and URLLC in some extent. Among three mission critical services namely eMBB, URLLC and mMTC, existing surveys have focused mainly on URLLC wireless networks in 5G and B5G. There is only one survey [15] that has covered the review of both URLLC and eMBB. However, [15] has reviewed the recent developments in URLLC and eMBB design based on B5G IIoT architecture. Moreover, it has explored the eMBB and URLLC studies in the literature where these services are provided in different wireless network architectures. Co-existence of both eMBB and URLLC service design is not addressed in any of the existing surveys. To fulfill this gap, we are providing a detailed survey on the eMBB and URLLC co-existence design in 6G networks. We primarily focus on various resource allocation techniques that have been adopted for the co-existence scenario. Main contributions of this survey paper are outlined below:

- We provide a comprehensive literature review on eMBB and URLLC co-existence in 6G wireless networks.
- We present various resource allocation techniques used for the eMBB and URLLC co-existence including scheduling, Transmit Time Interval (TTI), network slicing, distributed and federated learning, and RIS-UAV assisted resource allocation schemes.
- We present different applications of eMBB and URLLC system, and also provide insights on future research directions for the eMBB and URLLC co-existence.

C. PAPER STRUCTURE

Section II provides the basics of eMBB and URLLC co-existence system. Various resource allocation techniques for eMBB and URLLC co-existence are presented in section III. Section IV provides some applications of eMBB and URLLC services along with various use cases. Section V provides the future challenges with eMBB and URLLC co-existence, whereas the paper summary is provided in section VI. Figure 2 depicts the organization of the survey paper.

II. BASICS OF EMBB AND URLLC CO-EXISTENCE

In literature, different ways of allocating simultaneous eMBB and URLLC traffic have been discussed. Generally, four types of resource allocation schemes namely puncturing, superposition, Orthogonal Multiple Access (OMA) and Non-Orthogonal Multiple Access (NOMA) are widely studied in the literatures [18], [19], and [20]. The authors in [3], [21], and [22] have provided resource puncturing scheme for the simultaneous co-existence of eMBB and URLLC services. Providing URLLC traffic in the same network that already allocated resources to the eMBB traffic, leads to decrease the spectral efficiency of the eMBB traffic as the data and the transmission rates of eMBB traffic will be reduced. eMBB Quality-of-Service (QoS) is directly related to the allocated frequency resources to each eMBB user. This means QoS constraints of each eMBB user limit the maximum URLLC load. On the other hand, channel quality of simultaneous eMBB and URLLC services can directly impact the data rates of the eMBB users and latency and reliability of the URLLC traffic. Accordingly, less resources are needed for both services in the favorable channel conditions.

Three different types of models (linear, convex and threshold) have been studied in [18] for joint eMBB and URLLC scheduling problem. The authors have proposed a superposition/ puncturing scheme for rate loss for the eMBB service. A puncturing based scheduling technique has been proposed in [14] for maximizing rate utility. To determine the distribution of the resources for various users, a Hybrid Automatic Repeat Request (HARQ) technique has been used for latency and channel control [23]. The authors in [24] have proposed a risk sensitive approach for minimizing the rate loss using puncturing technique. A deep reinforcement learning (DRL) approach has been proposed to allocate the

TABLE 2. Overview of Existing Surveys on eMBB and URLLC.

Related Works	Topic	5G / 6G	eMBB or URLLC	eMBB /URLLC Co-existence	Resource Allocations
[13]	A Survey on Beyond 5G Network With the Advent of 6G: Architecture and Emerging Technologies.	✓	None of these		
[14]	A Survey of Scheduling in 5G URLLC and Outlook for Emerging 6G Systems.	✓	URLLC		✓
[15]	URLLC and eMBB in 5G Industrial IoT: A Survey.	✓	Both eMBB and URLLC	✓	
[16]		✓	URLLC		
[17]	5G Ultra-reliable Low-latency Communication Implementation Challenges and Operational Issues with IoT Devices.	✓	URLLC		
This work	Resource Allocation for Co-existence of eMBB and URLLC Services in 6G Wireless Networks: A Survey.	✓	Both eMBB and URLLC	✓	✓

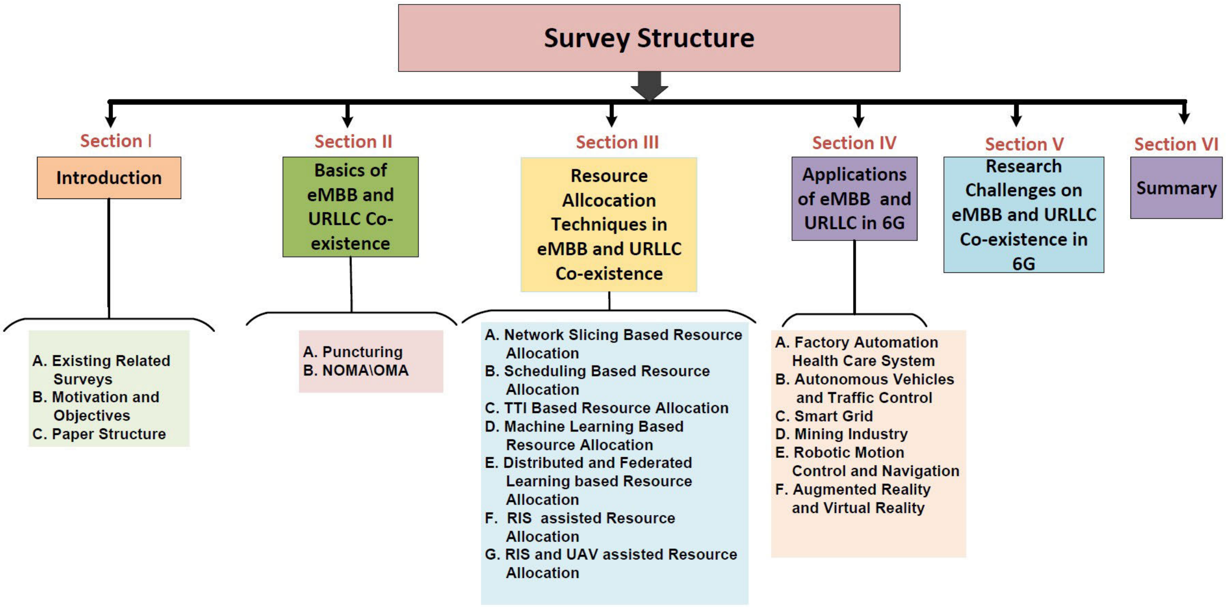

Survey Structure
Section I
Introduction
A. Existing Related Surveys
B. Motivation and Objectives
C. Paper Structure
Section II
Basics of eMBB and URLLC Co-existence
A. Puncturing
B. NOMA\OMA
Section III
Resource Allocation Techniques in eMBB and URLLC Co-existence
A. Network Slicing Based Resource Allocation
B. Scheduling Based Resource Allocation
C. TTI Based Resource Allocation
D. Machine Learning Based Resource Allocation
E. Distributed and Federated Learning based Resource Allocation
F. RIS assisted Resource Allocation
G. RIS and UAV assisted Resource Allocation
Section IV
Applications of eMBB and URLLC in 6G
A. Factory Automation Health Care System
B. Autonomous Vehicles and Traffic Control
C. Smart Grid
D. Mining Industry
E. Robotic Motion Control and Navigation
F. Augmented Reality and Virtual Reality
Section V
Research Challenges on eMBB and URLLC Co-existence in 6G
Section VI
Summary

FIGURE 2. Organization of the Paper.

URLLC service [25]. On the other hand, the authors in [26] have presented a superposition scheme for the allocation of URLLC traffic.

A. PUNCTURING SCHEME

Simultaneous eMBB and URLLC resource allocation in the same wireless network is enabled with the puncturing scheme [12]. In the puncturing scheme, some frequency resources are punctured by the URLLC traffic that have already been assigned to the eMBB users [12]. Such puncturing scheme has been shown in Figure 3. However, puncturing of such frequency allocations by the URLLC traffic reduces the allocated spectrum resources. These reduced spectrum resources ultimately decrease the data rate of the eMBB users [11]. Since the pre-allocated resource allocation directly relates to the minimum data rate requirements of the eMBB users, losing some resources to the URLLC traffic will degrade the QoS requirement of the eMBB users. Similarly, latency and reliability requirement of each URLLC user is limited by the ongoing QoS constraints of each eMBB user. In the superposition scheme, URLLC power allocation factor lies between 0.5 to 1 to accommodate the reliability of the traffic. Puncturing is the special case of superposition when the URLLC power allocation factor is 1. In the punctured mini-slot, base station allots zero power for eMBB user, and therefore, the interference cannot affect the URLLC traffic. There is no need for interference mitigation techniques to eliminate interference between eMBB and URLLC traffic in the given mini-slots. Rate loss of the eMBB traffic due to the puncturing resources by URLLC traffic can be modelled using different techniques. There are different rate loss models in the literature. Linear, convex and threshold model for the eMBB rate loss have been provided in [19]. In the linear model, it is assumed that the eMBB rate loss due to corresponding URLLC puncturing varies directly to the fraction of punctured mini-slots.

Different rate loss models for eMBB traffic due to puncturing scheme by URLLC transmission are illustrated in

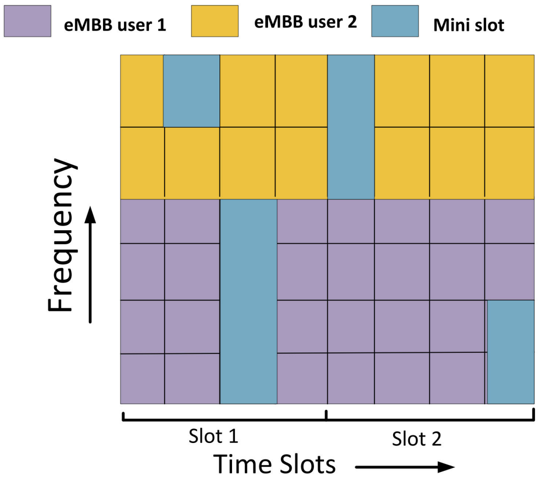


FIGURE 3. Puncturing technique for eMBB and URLLC co-existence.

Figure 4 as discussed in [19]. In this figure, y axis label h_u^s is defined as the rate loss function given by,

$$h_u^s(x) : [0, 1] \leftarrow [0, 1]. \tag{1}$$

$h_u^s(x)$ is actually defined in the equation

$$g_s^u\left(\psi_u^s, l_u^s\right) = r_u^s \psi_u^s \left(1 - h_u^s\left(\frac{L_u^s}{\psi_u^s}\right)\right). \tag{2}$$

where, $g_s^u(.., ..)$ is the rate allocation function that models the impact of URLLC on eMBB loss, r_u^s is the mean rate achieved by the user u in state s under the URLLC puncturing distribution, ψ_u^s is the total resource allocated to the user u in an eMBB slot in state s, L_u^s denotes the URLLC load superposed/punctured resource allocation of the user u in the mini-slot m when the channel is in state s. The function h_u^s captures the relative rate loss due to the URLLC overlap on the eMBB allocations.

In the linear rate loss model, $h_u^s(x) = x$. In the convex model, $h_u^s(.)$ behaves as a convex function as shown in Figure 4. In the threshold model, $h_u^s(x) = 1$ is an increasing function that shows the increasing behavior between 0 and 1.

These rate loss models will be briefly discussed below.

- Linear Rate Loss model: In linear rate loss model, fraction amount of punctured mini-slots by URLLC users has the direct impact on the eMBB rate loss. This means, the more punctured mini slots are, the more rate loss is and vice versa. For this model, there is a nice decomposition of the joint optimal scheduler. Arrival of URLLC packets in the mini slots of the eMBB slots can follow any distribution. For example, Poisson distribution [20], stochastic distribution [14], and binomial distribution [3] of URLLC packets arrivals are discussed in the respective papers. However, the non-liner utility functions and the time varying channel states can effect the URLLC signals to place in the eMBB mini-slots.
- Convex Model: In this model, rate loss is modeled by a convex function. For the convex model, finding the optimal solution is challenging since there is no structured decomposition property as in the linear model. A mini-slot homogeneous joint scheduling policy is adopted with the constant URLLC placement policy in the mini-slots inside an eMBB slot. In this model, a capacity region and concavity conditions are established to derive the effective eMBB rate after the punctured mini slots followed by an approximation algorithm for joint scheduling of both types of traffic. For such model, scheduling eMBB users share the bandwidth that slices across the frequency so that each user can utilize the entire slot duration to compensate for the eMBB rate loss by URLLC puncturing.

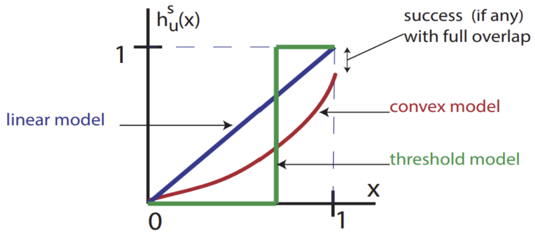


FIGURE 4. Rate loss function illustration in different models [19].

- Threshold Model: It is also known as 0-1 rate loss model because rate loss is assumed to be either 0 or 1 depending on the URLLC puncturing amount. In this model, a threshold is set up so that the eMBB traffic is unaffected unless the puncturing reaches the threshold, and the rate loss is assumed as 0 for this case. When puncturing reaches the threshold, then the rate loss is 1, i.e. a complete rate loss. A mini-slot homogeneous policy is assumed on the basis of rate proportional where the placement of the URLLC traffic in the mini-slots occurs in accordance with the eMBB resource allocations or eMBB loss threshold. This policy is advantageous in the sense that it reduces the eMBB probability loss in an eMBB slot.

B. NOMA/OMA SCHEME

In Orthogonal Multiple Access (OMA), radio resource allocations are orthogonal to time, frequency and code domain. Therefore, there is negligible interference among the users. OMA was mostly exploited to provide the wireless connectivity to the multiple users in the previous generations of the wireless communication systems. However, this scheme is considered insufficient for the upcoming 6G systems as this scheme is unable to provide the orthogonal resources to the maximum number of users simultaneously [21]. Therefore, OMA is not the good choice for providing resources to the large number of users. To provide the wireless connectivity to the increased number of users in more complicated networks scenarios with the fulfilment of the diverse requirement of ultra-reliability, high data rates, low latency and improved spectral efficiency, NOMA emerges as a new promising technique for upcoming 6G communications. Also, both OMA and NOMA are expected to utilize alternatively according to the service requirement based on the specific scenarios in upcoming 6G.

Multiple access to the resources are guaranteed through the power domain in NOMA system. This means that, in NOMA, multiple users are connected with various power levels unlike OMA. However, strong co-channel interference exists between multiple users in NOMA which needs to overcome using Successive Interference Cancellation (SIC). NOMA is basically an extension of the Superposition Coding (SC) applied in broadcast channels. Dealing with the NOMA over OMA in the co-existence of heterogeneous services such as eMBB and URLLC needs a development of the frameworks for both power control and scheduling policies. These frameworks can be utilized for modeling various scenarios related to the co-existence system such as uplink/downlink frame with traffic control model, overlapping eMBB/URLLC user transmission, eMBB/URLLC channel gains and interference noise mitigation models. In this regard, many research works have been done using both OMA and NOMA for the co-existence of eMBB and URLLC.

NOMA can be used as a competitive scheme to increase the efficiency of the resource utilization for eMBB and URLLC co-existence. Compared to OMA, NOMA performs better to improve the channel quality as it allows the superposition of the services even at the mini-slot level. This is achieved by adopting SC at the transmitter and SIC at the receiver. The use of Heterogeneous NOMA (H-NOMA) can be applied to the various problems such as clustering problem and minimization of resource conflicts in Layer 2 scheduling.

Employing NOMA technique encourages the simultaneous transmission of eMBB and URLLC packets in each mini-slot of the given slot in specified frequency block as shown in Figure 5(b). Therefore, existence of the interference is an unavoidable issue in NOMA based multiplexing technique. In the absence of the fixed channel quality, optimal SIC used in the usual case may not be employed for the strict latency-constraint situations. Therefore, use of SIC at the eMBB receiver based on the latency requirement of the URLLC user is a must [3]. Therefore, use of NOMA for eMBB-URLLC service applications is a difficult task. We may also argue that NOMA resource allocation contradicts the reliability and latency constraints as it should undergo some kind of interference mitigation techniques. The authors in [21] have utilized both OMA and NOMA multiplexing techniques to solve a power minimization problem. A Block Co-ordinate Descent (BCD) algorithm has been used where a look up table based approach is proposed for optimization of URLLC power consumption. In case of low average URLLC channel gain, NOMA results lower power consumption than OMA. Therefore, use of NOMA is beneficial to solve the energy efficiency problem in wireless networks.

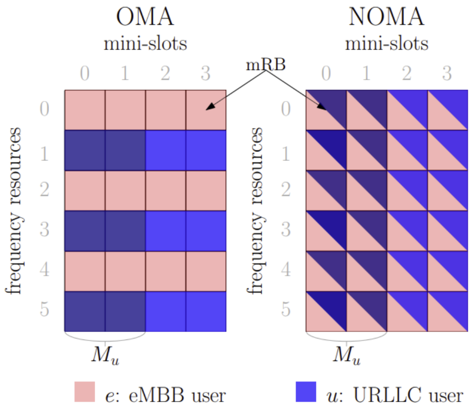


FIGURE 5. URLLC and eMBB traffic transmission in (a) OMA and (b) NOMA.

In case of OMA, an orthogonal allocation of the resources in the frequency domain is enforced to multiplex the users on different spectral resources. In such allocations, each mini-slot reserved for the URLLC packets can not be shared by the eMBB traffic [22]. One of the advantages of this scheme is that we do not have to consider interference among the URLLC and the eMBB traffic as these two traffic are transmitted in different slots as shown in Figure 5(a). However, spectral efficiency in OMA based multiplexing networks is lower than that in NOMA based networks. Therefore, use of NOMA can be an attractive solution for providing the heterogeneous services with higher efficiency of the networks and satisfying the stringent requirements of both types of the users. The use of NOMA can be benefited by employing the distributed and the federated learning in the eMBB and URLLC co-existence. In this regard, the authors in [10] have provided a NOMA-based distributed learning framework for the co-existence of the eMBB and the URLLC services in a Fog-radio architectures where both centralized and the decentralized learning approach have been adopted. Both low latency of the URLLC service and the high spectral efficiency for the eMBB traffic have been ensured by processing URLLC traffic at the Edge Nodes (ENs) and eMBB traffic at central processor in a Cloud-Radio Access Network (C-RAN). Distributed learning based co-existence for the eMBB and the URLLC has been discussed in section III-E. Different multiplexing techniques for eMBB and URLLC co-existence are provided in Table 3.

III. RESOURCE ALLOCATIONS FOR EMBB AND URLLC CO-EXISTENCE

Different multiplexing schemes for the co-existence of eMBB and URLLC have been proposed in the literature. Recent improvement in 3GPP standard supports multiplexing of two different types of services in 5G networks [2], [27]. In the superposition/puncturing scheme, time is divided into slots and each slot consists of several mini-slots to transmit URLLC packets upon their arrival during the resources occupied by ongoing service type transmissions [29]. In the superposition, BS allocates transmission power for both eMBB and URLLC traffic while in puncturing scheme, BS allocates zero transmission power for the eMBB traffic to allow the transmission of URLLC signals [3]. Some important works for the co-existence of eMBB and URLLC

TABLE 3. Multiplexing schemes on eMBB/URLLC co-existence.

Ref.	Objective	Resource Allocation Methods	Multiplexing Techniques
[3]	To minimize power consumption.	Look up table-based BCD technique.	OMA and NOMA
[22]	To maximize the utility for eMBB traffic.	Joint scheduling	Superposition
[23]	To Maximize rate utility.	HARQ	Puncturing
[24]	To Protect the eMBB users with low data rate.	Risk-sensitive based formulation.	Puncturing
[25]	To Maximize the eMBB data rate subject to a URLLC reliability constraint.	Decomposition and Relaxation based Resource Allocation (DRRA).	Puncturing/Orthogonal
[26]	To Maximize the eMBB network rate with respect to the URLLC QoS.	Contract based resource allocation.	Superposition

services along with different resource allocation strategies will be discussed in this section.

A spectrum partition method is applied to analytically investigate the eMBB and URLLC co-existence system [30]. A joint optimization problem is formulated to maximize the eMBB data rate while providing the URLLC to the users. A unique Data-driven Genetic Algorithm-based Spectrum Partition (DDGSP) is proposed to solve the proposed optimization problem. Simulation results show that proposed algorithm works better as both error rate and the computational complexity are improved compared to other existing works. In [31], a joint metric of minimum throughput of eMBB users and URLLC optimal demand for placement is evaluated to minimize the eMBB rate loss. The proposed metric helps to allocate the resource allocations for URLLC users in the mini slots. A joint resource allocation problem for the eMBB and URLLC co-existence has been formulated in [32] with the objective of satisfying URLLC interrupt probability and eMBB user rate simultaneously. The authors propose a Resource Block (RB) allocation scheme that provides the RBs for both types of users. The proposed scheme works in two stages: RB allocation scheme for eMBB users and URLLC transmission power calculation for ensuring URLLC reliability.

The authors in [33] have proposed a matching theory to find the suitable pair for the eMBB and URLLC users so that fairness among the eMBB users are established while maintaining the appropriate QoS parameters of the URLLC users. The objective of the work is to maximize the Minimum Expected Achieved Rate (MEAR) of eMBB users by satisfying the URLLC QoS constraints with respect to reliability and latency. Both eMBB and URLLC resource allocation problems are considered where the NOMA superposition-based technique is used for allocating URLLC traffic in mini slots of already transmitted eMBB traffic slots. The simulation results are compared with two other techniques namely contract-based and puncturing, and it has been shown that MEAR in this work is significantly improved compared to other two techniques. Besides, some other works based on the queuing techniques have also been studied in the literature. For example, a risk resistant scheme is proposed in [34] to allocate the resource allocations to satisfy the heterogeneous requirements of eMBB and URLLC co-existence. The objective of the work is to maximize the eMBB data rate with minimizing the URLLC threshold violation so that URLLC delay requirements will be maintained. Average URLLC packets delay is minimized by using a technique based on M/G/1 queuing model. Formulated non-convex optimization problem is transformed into convex optimization problem where the near-optimal solutions for the resource allocation are solved for each type of heterogeneous service. The authors have shown through the simulation that the fairness among eMBB throughput and eMBB users is improved by 30% - 34% when compared to other base-line techniques. The authors in [35] have proposed a Hybrid Multiple Access (HMA) solutions for addressing the issues in the co-existence of eMBB and URLLC. A spectral and energy efficient allocations of both traffic are scheduled using a Machine Learning (ML) based distributed hierarchical approach. Different case studies are presented to show the advantages of ML in allocating URLLC traffic in the case where there are limited radio resources. HMA-assisted with ML improves the data rate by more than 6 times with compared to the existing techniques. The scheduled URLLC traffic attains the maximum reliability even in the worst case with outage probability of 10^{-7}.

Medium Access Control (MAC) design is an important technique for scheduling the wireless signals of appropriate service-class such as eMBB and URLLC. Zaki-Hindi et al. in [36] have proposed a Medium Access Control (MAC) layer approach for up-link transmission of eMBB and URLLC traffic in an unlicensed spectrum for smart-factory scenario. The authors have designed a MAC model and have evaluated the reliability and delay for URLLC and throughput for eMBB users for low eMBB traffic loads. A successful transmission of the URLLC traffic is ensured by using a preemptive approach with the provision of high power at an instant when URLLC latency is close to delay constraint. Advantage of using this approach is that the impact on the eMBB rate loss is minimized with increasing URLLC performance. Mountaser et al. in [37] have worked on the Front Hauling (FH) networks where the multi-path diversity and erasure coding of the MAC frame is adopted for the improvement of reliability and latency of the URLLC users. The authors have investigated reliability and latency trade-off using a probabilistic model so that the reliable FH transport is possible by satisfying the average latency requirement. Numerical results prove the validity of analytical results and they ensure the successful transmission of eMBB and URLLC traffic. With the help

TABLE 4. Resource allocation techniques in eMBB/URLLC co-existence system.

Ref.	Objective	Resource Allocations Algorithms	Optimization Problem	Results
[10]	To reduce blockages and the collisions of the URLLC traffic by using eMBB spectral resources efficiently.	Heterogeneous-NOMA (H-NOMA) algorithm	Optimizing capacity of the digital fronthaul links connecting cloud and Edge Networks(ENs).	Use of fronthaul resources are minimized by using H-NOMA that are affected by URLLC interference by leveraging either puncturing or SIC at the ENs.
[30]	To maximize eMBB data rate while providing the strict URLLC to the users.	Data-driven genetic algorithm-based spectrum partition (DDGSP).	Multi-objective optimization (MOO) problem.	Reduces the system error rate from 48.2% to 4.1%.
[31]	To minimize eMBB rate loss with satisfying URLLC QoS.	Algorithm based on a joint Preference metric.	Optimally allocating resources to URLLC users problem.	Improves reliability of eMBB by 2.06% and 17.3% at low load and high load conditions respectively compared to random selection method.
[32]	To satisfy URLLC interrupt probability and eMBB user rate simultaneously.	Joint eMBB/URLLC resource allocation.	URLLC placement and eMBB scheduling	resource assignment for URLLC users guarantee URLLC QoS by achieving system's middle-end probability.
[33]	To use eMBB and URLLC preference profile to maximize the fairness among eMBB users while satisfying URLLC constraints.	One-one matching theory.	Maximize MEAR among eMBB users and maintain URLLC QoS.	Proposed technique outperforms the contract-based and puncturing techniques.
[34]	To maximize the eMBB data rate with minimizing the URLLC threshold violation.	Risk-resistant.	Minimize the average packet latency to realize the URLLC packet delay target.	Fairness among the eMBB throughput and the eMBB users is improved by 30% - 34%.
[35]	Radio resource management (RRM) for eMBB and URLLC co-existence.	Hybrid Multiple Access (HMA).	Spectral and energy efficient allocations of both traffic.	Data rate improved by more than 6 times with maximum reliability.
[36]	To enhance URLLC performance with minimum impact on eMBB performance.	Opportunistic preemptive approach.	Optimization for Power consumption minimization.	75% increase in eMBB throughput compared with conservative design with URLLC reliability of 99.99%.
[37]	To investigate reliability and latency trade-off to establish the reliable FH transport by satisfying the average latency requirement.	Multi-path diversity and erasure coding algorithm (MPC).	Erasure coding and multi-path transmission on the FH network.	An error probability of 10^{-5} is achieved by MPC with a low latency.
[38]	To maximize the overall system Energy Efficiency (EE).	Sliding Window (SW) based algorithm.	NP-hard EE optimization problem.	Reduction of power consumption by 16.7% and improvement of EE by 29.3% compared to other baseline systems.
[39]	To investigate an approach for satisfying QoS requirements of each user by using power allocations, bandwidth and punctured slices.	Lyapunov drift-plus-penalty (DPP) method	Problem for the minimization of the utility function of eMBB terminals tackling the with statistical delay constrains dealing with the probabilistic constraints.	Both statistical QoS of eMBB users and tight QoS user's requirements are satisfied with lower complexity than Genetic Algorithm (GA).
[40]	To minimize the total number of occupied RBS having fulfilled the QOS constraints requirement of both types of user.	Low complex heuristic algorithm.	Joint power and the Resource Block (RB) allocation problem as a mixed-integer programming problem.	Lower complexity of the proposed algorithm is achieved with respective to the exhaustive searching.
[41]	To investigate the performance trade-offs between URLLC and eMBB services under both OMA and NOMA, by considering puncturing, TIN, and SIC.	Orthogonal (OMA) and Non-Orthogonal Multiple Access (NOMA)	C-RAN theoretic optimization	TIN always outperforms puncturing in analog C-RAN while NOMA with SIC performs the best among all techniques.
[42]	Network sum rate maximization with meeting the requirements of URLLC latency and eMBB minimum rate constraints.	Heuristic scheduling algorithm.	Mixed-integer non-linear programming.	Latency of the URLLC users is improved in heuristic algorithm than optimization algorithm.
[43]	eMBB throughput enhancement and improvement of URLLC delay.	A dynamic multi connectivity (MC)-based joint scheduling algorithm.	Two-step optimization: frequency allocation optimization and decomposition to a difference of convex (D.C.) functions about power.	A maximum of 49% increase in the throughput.
[44]	To maximize fair eMBB data rates with satisfying their minimum data rate requirements.	QoSG-RA algorithm	Optimization of network resource allocation.	QoSG-RA meets QoS requirements of both types of users by offline approach and maximizes fairness of the eMBB data rates via online approach.

of both analytical and the numerical procedure, The authors compare FH resource allocation for OMA and NOMA cases and found that proposed FH technique has better performance compared to both OMA and NOMA.

The authors in [38] have proposed a Sliding Window (SW) based algorithm to maximize the system Energy Efficiency (EE) in eMBB and URLLC co-existence system. The proposed technique attempts to find the resource pool for each type of services and then allocates the appropriate Resource Blocks (RBs) by using a linear programming relaxation approach. The proposed SW technique has low complexity to solve the EE optimization problem. Scalability of the proposed technique is demonstrated regarding to the grid sizes and maximum transmit power. The superiority of the proposed SW technique is demonstrated by numerical results with the reduction of power consumption by 16.7% and improvement in EE by about 29.3% compared to traditional approaches. However, energy efficiency of the considered wireless network framework also depends on the power allocation strategies along with RBs management. Therefore, a joint allocations of power, bandwidth and resource blocks need to be optimized for designing energy efficient wireless networks for eMBB and URLLC co-existence. To this end, a joint optimization problem for bandwidth and power allocations for the co-existence of heterogeneous services has been provided in [39]. The authors have investigated an approach for satisfying QoS requirements of each user by using power allocations, bandwidth and punctured slices. To meet the balance between the long term URLLC constraints and short term optimization problem, Lyapunov Drift-Plus-Penalty (DPP) method is used where the short term optimization problem is constantly utilized to satisfy the long term constraints. Furthermore, a well-known Block Co-ordinate Descent (BCD) algorithm is adopted where the main problem is divided into two sub-problems in each block and the optimized parameters in each block is used for the remaining block so that overall computational complexity of the proposed solution is reduced. The authors then use online matching theory for solving the integer programming in allocations of RBs punctured slices. Thanks to the adopted dynamic resource allocation and puncturing strategy (DRAPS), eMBB and URLLC scheduling problems are solved addressing the issues of dynamic CSI and random URLLC packet arrival. The authors in [40] have formulated a joint power and Resource Block (RB) allocation problem as a mixed-integer programming problem. Objective of the formulated problem is to minimize the total number of occupied RBs having fulfilled the QoS constraint requirements of both types of users. Intuitively, formulated problem solves the amount of required power and required RBs to be allocated to each type of users satisfying their service requirements. A low complex alternating optimization algorithm is applied to obtain the global optimal solution. Lower complexity of the proposed algorithm with respective to exhaustive search is verified by an extensive simulation.

Apart from the above mentioned resource allocation techniques, some other works in the literature have studied eMBB and URLLC co-existence for Radio Access Network (RAN). Both distributed and centralized schemes have been used in RAN networks to provide specific services classes such as eMBB and URLLC. In [10], a multi-cell F-RAN architecture is studied for the eMBB and URLLC co-existence. A non-orthogonal radio resources are shared between the eMBB and the URLLC users where eMBB and URLLC traffic are processed at the edge and the cloud respectively. Even in presence of inter-service interference, NOMA provides better resource allocations than in OMA case. Moreover, the proposed H-NOMA resource allocation technique offers higher spectral efficiency for the eMBB traffic with low latency for URLLC access. The proposed approach also results small number of collisions and blockage for URLLC transmission. Advantages of H-NOMA depends on how URLLC interference is managed on eMBB signals in case of small front hauling capacity. A SIC technique is used in the up-link transmission prior to the front-haul compression. Another work in [41] provides a analytical study on the eMBB and URLLC co-existence problem for up-link Cloud Radio Access Network (C-RAN) scenario. Both OMA and NOMA based radio resource slicing techniques are adopted to provide the performance of both types of heterogeneous services with information-theoretic view. Both additive Gaussian noise and the inter-cell interference are considered for all the three techniques namely SIC, Treating Interference as Noise (TIN) and puncturing. From the simulation results, it has been shown that NOMA technique provides improved eMBB data rate compared with OMA in analog front-hauling C-RAN architecture ensuring reliable URLLC transmission with minimum latency. Furthermore, TIN seems to perform better than puncturing whereas NOMA with SIC outperforms all the considered techniques. The authors in [42] have proposed a sum rate maximization problem using a RAN. Then, a resource allocation problem is formulated by considering Adaptive Modulation and Coding (AMC) subject to various constraints such as minimum rate constraint, latency related constraint and service isolation constraints. As the formulated problem is NP-hard and non-tractable involving binary variable and AMC, the authors use non-linear approximation to relax mathematical intractability and use penalized reformulation to convert binary constraint to a box constraint. As a result, the associated non-optimization problem is converted to bi-convex problem, and a sub-optimal solution of original optimization problem is obtained.

Some authors have also explored the scheduling techniques for the co-existence of eMBB and URLLC. These works have utilized low complexity heuristic scheduling schemes to reduce the complexity of eMBB and URLLC scheduling problems. Zhang et al. in [43] have worked on a joint scheduling framework based on dynamic multi-connectivity (MC) for eMBB and URLLC system. Queuing of URLLC can be avoided with eMBB, and URLLC are sliced with each other. Here, a modified Effective Capacity

(EC) technique is proposed to evaluate the performance of the co-existence framework. The objective of using EC model is to guarantee the URLLC QoS without decreasing eMBB throughput. A two-step optimization technique is used where the MINLP in EC model is converted into internal non -linear programming. Compared to traditional EC framework, proposed EC framework performs better as both system throughput and URLLC latency are improved. The authors in [44] have proposed a new scheduling approach called QOSG-RA for eMBB and URLLC resource allocations. The proposed QOSG-RA utilizes the hybrid on-line and off-line resource allocations. Off-line resource allocation scheme is performed to guarantee the QoS of both eMBB and URLLC service requirements while on-line resource allocation is used to maximize the eMBB data rate depending on run-time knowledge. The simulation work shows that the proposed QOSG-RA approach is effective on achieving the eMBB data rate requirements with QoS while satisfying the URLLC latency and reliability requirements, and outperforms the other state-of-art techniques. A summary of different resource allocation techniques adopted for the co-existence of eMBB and URLLC services along with the formulated optimization problems and outputs of these techniques are presented in Table 4. A detailed study on different types of resource allocation techniques that have been adopted for the co-existence of eMBB and URLLC have been presented in the following sections. For each section, we draw a short conclusion as lesson learned that summaries the important aspects of the research works adopting respective resource allocation techniques.

A. NETWORKING SLICING BASED RESOURCE ALLOCATION

Co-existence of the mission critical services such as eMBB and URLLC can be achieved by accommodating network slices. Through the network slicing, a design of customized slices of network is provided to guarantee the various requirements of service heterogeneity. Specially, network slicing provides a sub-optimal use of the networks which helps to fulfil the requirements regarding latency, reliability and QoS of both eMBB and URLLC services. Various network slicing techniques have been used in the literature for the co-existence of eMBB and URLLC services design. Orthogonal and Non-orthogonal networks slicing are two important techniques that have been widely used in various works.

A dynamic resource allocation scheme with the combination of the scheduling and the network slicing is provided in [26]. To minimize the impact on ongoing eMBB service due to URLLC users and provide spectrum access to the URLLC users while reducing the intervention on incumbent eMBB users, the authors have used the contract theory framework by utilizing the network slices. In [45], a saddle-point approximation technique is provided to unify the information-theoretic framework for blending an infinite and finite block length analysis respectively. Puncturing and superposition coding are used for providing down-link co-existence strategies. Both puncturing and the superposition techniques are adopted to evaluate the performance of eMBB and URLLC services. The simulation results show that the superposition coding is superior than puncturing technique in terms of spectral efficiency for the eMBB users and target reliability for the URLLC users. However, the URLLC performance could be enhanced with the puncturing scheme in a situation when the multi-user interference exists.

Spectrum utilization is a key aspect of fulfilling throughput requirement for eMBB users while guaranteeing the URLLC QoS. To this end, the authors in [46] have proposed a matching algorithm with minimum complexity to improve the spectrum utilization. A Multi Users Multiple Input Multiple Output (MU-MIMO) technology has been adopted to allocate the simultaneous transmission of eMBB and URLLC traffic. Use of matching algorithm has promoted an eligible replacement of the eMBB users by the URLLC users to ensure the transmission URLLC traffic. The simulation results show that the proposed technique achieves better performance with lower complexity compared to exhaustive search method.

In [47], the authors have used Rate Splitting Multiple Access (RSMA) to split the URLLC message into two sub-messages with different power allocation factors. Two sub-messages can be recovered using SIC at BS. Comparing with OMA and NOMA, RSMA has shown better performance in terms of sum-rate and reliability. Thus, an adjustment in rate splitting factor results URLLC sum-rate improvement based on average SNR rather than instantaneous Channel State Information (CSI). A eMBB-URLLC multiplexing technique for a downlink transmission scenario is proposed in [48] with an objective of reducing the size of punctured eMBB symbols by exploiting possible similarities between two types of symbols. For this, BS only scans the eMBB traffic symbols which are punctured by URLLC users that have maximum symbol similarity. The idea is to use the symbol region similarity to accommodate the various constellations of eMBB and URLLC traffic. The authors propose an analytical scheme by deriving a closed-form expression for Symbol Error Rate (SER). What it has been demonstrated from the proposed puncturing scheme is that there is an enhancement in system information rate when the URLLC load is twice with the same SER and an achievement of gain up to 10 dB with respect to baselines techniques.

A concept of network softwarization has been used in [49]. Both Software Defined Networking (SDN) and Network Functions Virtualization (NFV) are implemented for sharing a common infrastructure of eMBB and URLLC services. A resource sharing algorithm based on two-level MAC scheduling has been leveraged for computing and adjusting the required radio resources for each eMBB and URLLC service. The simulation results show that the proposed network slicing technique satisfies the heterogeneous requirements of both types of network slices. In [50], a mixed integer non-linear program is proposed for the resource allocations with the network slicing. A chance constraint is used to isolate eMBB and URLLC networks slicing and traffic load uncertainty where a sub-optimal solution of formulated problem has been determined by using penalized SCA technique. It has been shown that the proposed algorithm has a lower complexity and better performance in terms of throughput and reliability than the baseline approaches. In [51], a Rate-Splitting Multiple Access (RSMA) approach is proposed for the analysis of various slicing schemes in an up-link scenario. The authors show that the RSMA is superior than OMA and NOMA in terms of the achievable rate region. The RSMA has advantages of increasing data rate of users for different purposes. Also, the RSMA reduces the computational complexity by making flexibility in decoding two split streams of each user. Thus, splitting message depending on the service requirement with an allowance of appropriate decoding order design is possible. A brief summary of the implemented network slicing techniques has been provided in Table 5.

a: LESSONS LEARNED

Proper management and accommodation of network slicing can be an effective technique to deal with the heterogeneity of network and its requirement to provide various service classes within the network. For the eMBB and URLLC co-existence models, three main network slicing techniques have been utilized in various works in the literature: OMA, NOMA and RSMA. Works in [26], [45], and [46] have demonstrated use of OMA and NOMA by adopting both puncturing and superposition schemes whereas [47], [48], and [49] have used the RSMA as a superior network slicing technique than OMA and NOMA. Orthogonalization and decoding interference are the traditional techniques to mitigate the interference in OMA and NOMA, respectively. Besides these, RS is able to mitigate interference by partially decoding interference and partially treating interference as noise by splitting the given message into common and private parts. Reliability and low latency in URLLC are ensured in RSMA as it facilitates robustness to inaccurate CSI. As a result, RSMA has been proven to have unique benefits such as computation, energy and spectral efficiency.

B. SCHEDULING BASED RESOURCE ALLOCATION

In [52], the authors present a scheduling technique for the co-existence of eMBB and URLLC scenario. A spatial preemptive scheduler based on the null-space is proposed for the large number of users. A cross objective optimization problem for the guarantee of URLLC QoS and extracting maximum possible eMBB ergodic capacity has been formulated based on proposed scheduler framework. In the proposed scheduler framework, critical URLLC traffic is scheduled without any interference due to utilization of system spatial degree of freedom. As a result, URLLC decoding ability is enhanced without much impacting the eMBB performance. From the simulation results, it is shown that the proposed technique outperforms the state-of-arts techniques in terms of the URLLC latency minimization and ergodic eMBB capacity improvement.

The authors in [53] have proposed a puncturing based co-scheduling problem of eMBB and URLLC traffic. Objective of this work is to minimize the Minimum Expected Achieved Rate (MEAR) of eMBB traffic satisfying URLLC QoS. The original problem is divided into two sub problems. One is eMBB and URLLC user scheduling and the other is to maximize MEAR of eMBB user. The authors use a Penalty-based Successive Upper-bound Minimization (PSUM) to solve the scheduling issues of the eMBB users whereas an Optimal Transportation Model (OTM) is used for solving the same problem of URLLC users. Moreover, an heuristic approach with lower complexity is also provided to solve the scheduling problem.

In [54], an Enhanced Null Space-based Preemptive Scheduler (eNSBPS) is proposed to jointly optimize the URLLC and eMBB traffic in the networks. With the help of eMBB subspace projection, proposed technique ensures the instant sporadic URLLC transmission without queuing. As a result, an extremely reliable URLLC transmission is guaranteed. On the other hand, eMBB achievable capacity is maximized as the lost or affected eMBB capacity that can be recovered through aligning subspace of the victimized eMBB users. The proposed scheduling technique has shown better performance than other scheduling schemes in terms of the URLLC latency target and improvement in overall cell spectral efficiency. Another work in [55] introduces a joint link adaption and resource allocation policy. Use of the proposed technique is able to dynamically adjust the block error probability of URLLC transmission according to given load per cell. Extensive simulation results show that the proposed technique is extremely useful regarding the URLLC latency. Compared to the conventional state-of-arts techniques, the proposed technique reduces the URLLC latency from 1.3 ms to 1 ms at 99.999% percentile. The results also show that there is only 10% degradation on the eMBB throughput in comparison to conventional scheduling policies. The simulation results cover a sensitivity analysis for determining the performance analysis of both eMBB and URLLC services for various loads. These results are helpful for analyzing the relation between maximum URLLC loads and to satisfy the corresponding URLLC requirements.

The authors in [56] have used a NOMA based puncturing technique in MIMO for solving a co-scheduling or co-existence problem of eMBB and URLLC. A joint power allocation and user selection technique are used to formulate the objective function of maximizing eMBB data rate with satisfying the QoS and latency requirements for the URLLC users. A clustering mechanism for eMBB users is introduced to trade-off between system throughput and computational complexity. Then, two sub problems namely scheduling of user selection and power allocations are introduced to replace the original problem. User selection problem is solved using

TABLE 5. Research on network slicing techniques for eMBB and URLLC co-existence scenario.

Ref.	Objective	Network Slicing Techniques	Optimized Parameters
[45]	To reduce size of punctured eMBB symbols.	Puncturing and superposition coding.	Average SE per user and network availability.
[46]	To improve spectrum utilization to enhance throughput of eMBB QoS of URLLC.	Matching algorithm	Eligible replacement of eMBB users by URLLC users.
[47]	To improve sum-rate based on average SNR.	Rate Splitting Multiple Access (RSMA)	Rate splitting factor
[48]	To reduce size of punctured eMBB symbols.	Symbol region similarity	Symbol Error Rate (SER).
[49]	To slice eMBB and URLLC frequency and to isolate RAN resources.	Network Softwarization.	Radio resources to be used by each deployed network slice.
[50]	To guarantee the load demand for each eMBB and URLLC user with the QoS constraints.	Chance constraint technique.	Allocation of Remote Radio Heads (RRHs) and Physical Resource Blocks (PRBs).
[51]	To achieve a larger rate region.	RSMA	Decoding rate of different class service.

Gala-Shapely (GS) method while SCA is adopted for power allocation sub-problem. In order to find the global solution, an iterative algorithm with low complexity is used. In another work [57], a cancellation mechanism is introduced that allows the URLLC user to preempt the eMBB transmission. The authors have used a 3GPP attack model to prove that stringent QoS requirement for the URLLC user may be a threat for interference of both eMBB and URLLC traffic. A system level simulation is conducted to evaluate and investigate the proposed attack model. Through the simulation, it has been shown that the amplification on overall impact on both eMBB throughput and the URLLC latency is enhanced for small number of users when the attackers leverage synchronization among the compromised URLLC users. Table 6 presents the outlines of some scheduling based resource allocation schemes with their contributions to eMBB-URLLC co-existence scenario.

a: LESSONS LEARNED

Different scheduling techniques are useful to spontaneously schedule the URLLC service upon their arrival. Such scheduling techniques have been presented in [52], [53], [54], [55], [56], and [57]. The works [53] and [54] emphasized on the QoS of URLLC users without impacting the eMBB throughput whereas [54] deals more on URLLC reliability than the eMBB data rate using an effective scheduling technique namely eNSBPS which outperforms the other schedule-based techniques by altering the system optimization to a region where the URLLC QoS is instantly guaranteed. One of the important conclusions revealed from the scheduling of eMBB and URLLC services is that the URLLC QoS and the reliability depend on the available URLLC loads [55] i.e. minimum URLLC loads result better performance than the higher URLLC loads to maintain the same eMBB throughput loss.

C. TRANSMIT TIME INTERVAL (TTI) BASED RESOURCE ALLOCATION

In literature, many works have already been progressed for the co-existence of the eMBB and the URLLC in B5G and 6G using different scheduling techniques. Among these scheduling techniques, Transmit Time Interval (TTI) management is the one through which different class of services are scheduled within the same wireless networks. The length and the frequency of the flexible TTI are determined by considering appropriate RBs. The basic idea of using RBs with different TTIs is to provide the transmission of different service-class as shown in Figure 6 where one RB is represented by each square grid that comprising of certain time axis and frequency axis. In the figure, all available resources are represented by the entire area. With the provision of TTI based heterogeneous service provision, usually large TTIs are used for the eMBB users whereas the short TTIs are used for the URLLC service transmission. URLLC service transmission interrupts already assigned eMBB transmission as some code blocks are replaced by URLLC users to guarantee URLLC latency requirement. This process may lead to the reduction of eMBB QoS and data rate. Flexible TTIs and flexible frame structure may solve this issue.

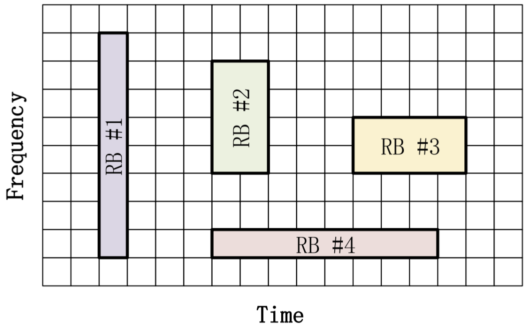


FIGURE 6. Use of flexible TTI for transmitting different class of services [63].

In [58], the authors provide a framework of multiplexing eMBB and URLLC services with the help of Neural Network (NN). Flexible frame structures for both services are generated by using Adaptive Neuro-Fuzzy Inference System (ANFIS) based approach where the data obtained from FIS

TABLE 6. Different scheduling techniques for eMBB and URLLC co-existence scenario.

Ref.	Objective	Scheduling Technique	Contribution	Evaluation
[52]	To minimize the impact on overall ergodic capacity and to guarantee an instant scheduling for sporadic URLLC traffic.	Null-space-based preemptive scheduler.	NSBPS scheduler safeguards the URLLC traffic from potential inter-user interference by enforcing sufficient spatial separation through subspace projection.	URLLC outage latency of 1 ms and average cell throughput of 37.5 Mbps.
[53]	To minimize MEAR of eMBB traffic satisfying URLLC QoS.	Puncturing based co-scheduling.	This technique uses an algorithm based upon Minimum Cell Cost (MCC) and Modified Distribution (MODI).	A minimum MEAR value of 18.0 Mbps with a probability 0.231 is achieved.
[54]	To alter the system optimization to instantly guarantee URLLC QoS and to recover delay-tolerant eMBB QoS.	Enhanced spatial preemptive scheduling.	The proposed technique ensures a quick signal subspace free of interference thus guaranteeing the extreme low latency for URLLC eMBB performance.	Overall cell spectral efficiency is improved with an average gain achievement 3.2 dB in eMBB post-detection carrier-to-interference-ratio.
[55]	To adjust block error probability of URLLC small payload transmissions according to the instantaneous experienced load per cell.	Joint link adaptation and resource allocation policy.	This technique helps for sensitivity analysis to determine the performance analysis of both eMBB and URLLC services for various loads.	Proposed technique reduces the URLLC latency from 1.3 ms to 1 ms at 99.999% percentile. There is only 10% degradation on eMBB throughput.
[56]	To maximize data rate of eMBB users while satisfying the latency requirements of URLLC users.	Gale-Shapley (GS) matching algorithm.	GS matching algorithm provides a unique technique for eMBB and URLLC user selection in two distinct sets based on player's individual preference.	Total eMBB throughput is 35 bits/s/hz where as total eMBB throughput per mini slot is 22.5 bits/s/hz.
[57]	To amplify degradation of eMBB throughput.	DDoS attack mechanism.	A cancellation mechanism is introduced that allows URLLC user to preempt eMBB transmission.	URLLC latency is no more than 2.5 ms with a reliability of 0.999.

is used for both training and testing of the NN model. The simulation results show that the co-existence problem is solved by the flexible frame structure that ensures the QoS of both service classes with much short delay for URLLC users. In another work [59], latency performance in a multi-user cellular networks is designed for downlink communication. The authors use the flexible frame structure for configuring flexible TTI size per user basis for each type of service requirement. From the simulation results, it is shown that use of short TTI is beneficial to achieve latency requirement of URLLC users. However, low URLLC loads are preferable for implementing short TTI for URLLC achievement. For a higher load, the performance is improved by using a longer TTI configurations with lower relative control overhead and higher spectral efficiency.

In [60], a user-centric scheduling approach is proposed that exploits the flexible TTI and dynamic Time Division Duplex (TDD). In this work, traffic load is configured by adapting the down-link to up-link ratio and TTI duration. A real user-centric scheduling approach is implemented by solving a joint optimization problem that is formulated based on the individual requirement of each user in terms of latency and throughput. The simulation results show that the proposed technique works better than the traditional technique based on dynamic TDD and fixed TTI in terms of both eMBB throughput and URLLC latency requirement. The authors in [61] have proposed a flexible Frequency Division Duplex (FDD) to optimize the resource allocations per link basis. In that work, data rate for each individual user is managed by using in-resource physical layer control signaling. The advantage of using this technique is to offer a short round trip for air-interface resulting to maximize the throughput thus guaranteeing efficient support for machine type communication.

The authors in [62] have proposed a MAC layer framework in mmwave band that includes the enhancements in flexible transmission times along with the ability of multiplexing directional control signals. The relation between control overhead and its utilization with number of parameters such as periodicity, user numbers, SNR and antenna gains have been analytically derived. The analytical derivation covers and incorporates various critical features of the mmwave transmission. Analysis and the simulation results show that the proposed flexible frame structure provides various important benefits than the fixed frame structure. Additionally, a lower overhead can be obtained with fully digital beamforming structure than the analog structure under the identical power conditions. In [63], a shortened TTI mechanism is implemented for downlink system. In that work, a flexible TTI and traffic arrival rate are combined to schedule the URLLC traffic. The authors propose an algorithm that takes care of impact of shortened TTI on down-link communication. Although the proposed algorithm is able to improve throughput and lowers the URLLC latency, it works only for small URLLC loads. As URLLC load increases, the performance of the algorithm also degrades, and the algorithm fails to work for extremely high loads. This is because, at higher loads, relative control overhead is more than the queuing delay. In such case, length of the TTI needs to be increased to reduce the impact of the queuing delay. Research on different TTI based resource allocation techniques have been concisely presented in Table 7.

a: LESSONS LEARNED

Resource allocations in eMBB and URLLC co-existence can be facilitated by using different TTI for service specific transmission. For example, short TTI is preferred for URLLC traffic and long TTI is beneficial for eMBB traffic. The concept of varying length of TTI reduces the impact of eMBB rate loss due to URLLC transmission. Aforementioned works in [58], [59], [60], [61], [62], and [63] utilize the flexible TTI for the transmission of eMBB and URLLC traffic. Flexible TTI has been designed for the downlink transmission in [59] and [63] whereas [58], [61], [62], and [64] deal with both uplink and downlink transmissions. The general idea of providing short TTI for URLLC and long TTI for eMBB traffic is not always true. The effectiveness of length of TTI depends on the number of URLLC loads to be transmitted. In case of large number of URLLC loads, the length of TTI needs to be increased to accommodate the relative control overhead [63] which is more than the queuing delay. Furthermore, both dynamic TDD and FDD techniques can be employed with the flexible TTI. FDD can use separate frequency bands for uplink and downlink communications [63]. However, TDD utilizes the same frequency band for uplink and downlink but the transmission is limited to only one direction at a time [64].

D. MACHINE LEARNING (ML) BASED RESOURCE ALLOCATION

The authors in [64] propose a novel technique based on self-adaptive flexible TTI where a ML technique is used to design the flexible scheduling for eMBB and URLLC. An appropriate selection of TTI per user is accomplished by designing a decision algorithm based on Random Forest (RF) which is named as RF-ETDA. As TTI is a foundation for the resource allocation and scheduling, there is an improvement of average URLLC delay by 45.64% and average Packet Loss Rate (PLR) is improved by 59.17%. Since eMBB service is not sensitive with latency, this class of service prefers larger TTI than the URLLC class users. Alsenwi et al. in [65] have proposed Deep Reinforcement Learning (DRL) based framework to formulate an optimization problem aiming at maximizing eMBB data rate and minimizing URLLC latency. The formulated problem has the constraints of URLLC latency requirements and eMBB variance in data rate so that the impact on the eMBB service can be minimized. The considered problem is divided into two parts namely resource allocations for eMBB and scheduling for URLLC. Also, the optimization problem is divided into different sub-parts, and sub-optimal solution is approximated in each part by converting the original non-convex problem into locally convex problem. From the simulation results what is proved that the proposed DRL based scheduling scheme provides the eMBB reliability of higher than 90% while satisfying the stringent URLLC latency requirement.

Tang et al. in [66] have solved an optimization problem relating to the QoS of heterogeneous class of eMBB and URLLC services. The authors in this work have provided the flexible numerology structure defining the flexible TTI satisfying the QoS of both services. The proposed optimization problem is NP-hard and difficult to solve. Hence, the authors use Markov Decision Process (MDP) based DRL technique with the purpose of dynamic allocation of the resources. In their extensive simulation works, the authors compare the results of the proposed technique with some other well-known strategies namely random, greedy and sequence techniques. It has been concluded that the proposed MDP-DRL based strategy uses an innovative joint scheduling strategy DRSA that outperforms all the above strategies. Evidences from the simulation works prove that the URLLC service's loss rate is reduced by 53.8%, 28.6.% and 43.7% in DRSA than random, greedy and sequence strategies, respectively. Additionally, average throughput of the DRSA is higher than the above strategies by 23.9%, 14.8% and 7.1%, respectively.

The authors in [67] have introduced RAN resource slicing based on the user intent for the eMBB and URLLC co-existence. The objective is to maximize the service level agreement (SLA) satisfaction degree of various types of users. The authors have proposed an intent-driven Multi-agent DRL (MA-DRL) based algorithm. The matching degree between the practical performance and the user intent has been evaluated by designing a reward function using MA-DRL technique. The simulation results have shown that the proposed technique outperforms the existing benchmark techniques. Naveen Kumar and Ahmad in [68] have proposed a Neural Network (NN) based technique for the prediction of network slices in co-existence of the eMBB and URLLC services. A dynamic network slicing is required for the various channel conditions and heterogeneous QoS requirement for different users. Advantages of using proposed NN technique for the network slicing is to help the management of slices in accordance with per-user throughput requirement. Also, the proposed NN technique helps to dynamically improve the feature extraction. The authors show, by the extensive simulation, that the proposed NN based slicing management technique provides better results in terms of utilising the resources and satisfying the QoS of each class of users simultaneously. In [69], a Deep Deterministic Policy Gradient (DDPG) algorithm is proposed with the objective of maintaining a trade-off between the heterogeneous service requirements for eMBB and URLLC. DRL based DDPG technique jointly optimizes overlapped URLLC positions and bandwidth allocations with the help of observation in channel variations and arrival of URLLC traffic. Employed DDPG technique achieves the good trade-off between eMBB and URLLC service requirement without degrading eMBB QoS.

TABLE 7. Research on Transmit Time Interval (TTI) based co-existence models for eMBB and URLLC services.

Ref.	Objective	TTI based Techniques	Achievement
[58]	To ensure QoS of both service class users with the guarantee of short delay for URLLC users.	ANFIS based approach.	Guarntee of QoS for both service class users with much short delay for URLLC users.
[59]	To analyze trade-off between queuing delay and TTI size on a system level.	OFDMA-based frame structure.	Short TTI of 0.25 ms is an attractive solution for URLLC.
[60]	To meet individual requirements of each user in terms of latency and throughput.	User-centric scheduling approach.	URLLC latency of below 2 ms for supporting 12 Mbps eMBB throughput for more than 300 packets/s.
[61]	To optimize resource allocations per link.	FDD	Offers a short round trip for air interface resulting to maximize throughput.
[62]	System evaluation for URLLC traffic, SNR and control periodicity using various statistical models.	mmWave MAC layer frame structure.	Low latency can be achieved by scheduling the granular positioned control channel within the frame.
[63]	To take impact of shortened TTI on throughput and latency.	Downlink Scheduling Alogrithm.	URLLC loads and TTI length vary inverse to get high throughput and the low latency.

Zhang et al.in [70] have proposed a scheduling problem based on stochastic optimization with an objective of maximizing the total eMBB throughput and URLLC utility. Since the mathematical expression for analyzing the expected value of URLLC served packets is in the closed form, the authors used the Bayesian optimization technique combined with the additive structure algorithm to solve highly computationally complex problem for the steady state URLLC queue solution. The objective of using hybrid optimization technique is to maintain the normalized eMBB throughput and to decrease the dropping probability of URLLC packets by meeting the dalay requirements. The simulation results show that the proposed Bayesian algorithm has performed better than the random search in terms of finding the optimal solution. Comparing to the other benchmarks techniques, the proposed queuing scheme and the URLLC scheduling policy have provided higher throughput with the target URLLC packets transmission.

A water-filling algorithm [71] is used to allocate the eMBB resource in [72]. A DRL based physical layer resource slicing scheme between the eMBB and URLLC traffic is adopted. In that work, a DRL model is trained so that the agent decides the allocation of URLLC traffic inside the fully occupied time-frequency grid by eMBB users. The dynamic puncturing of eMBB traffic by URLLC users occurs in the mini slots. DRL agent in the proposed work balances the minimum eMBB code-word in outage by maintaining the URLLC latency requirements. Channel State Information (CSI) at BS is assumed to be perfectly known thanks to known sufficient time available to exchange Channel Quality Information (CQI).

A unique network slicing scheme using DRL based framework is provided by Suh et al. in [73]. They have used a DRL-NS algorithm with an objective to improve the overall data rate requirements while satisfying the latency requirements. An action elimination method is employed to reduce the action space so that the decisions for the unwanted resource allocation that impact the latency requirements are eliminated. As a result, optimal resource allocations policy can be achieved by learning improvement using the proposed model. The proposed technique outperforms the conventional algorithms meeting the throughput and the QoS requirements. The proposed DRL-NS model can be extended to solve many other issues including beam management and link scheduling management for cognitive radio resources. From the simulation results, it has been shown that there are 19% and 27% data rate improvements in comparison to regression-tree based slicing technique and equal allocation technique respectively at the maximum delay of 1 ms. The computational complexity of the proposed technique is low as the convergence time for the algorithm at 20 dB SNR level is only 1730 s.

In [74], the authors have proposed a DRL based algorithm aiming to balance the Key Performance Indicators (KPIs) for the resource allocations in different heterogeneous service including URLLC and eMBB. In this work, latency and reliability for URLLC services and throughput requirements of eMBB users are jointly optimized by utilizing time flexibility of time-frequency grid. As a result, both resource allocations and power allocations are jointly performed. The performance of the proposed DRL based algorithm is compared with a priority-based proportional fairness algorithm that utilizes the fixed power allocations. There is 29% improvement in power allocations and 21 *times* improvement in the throughput. A brief summary on resource allocation for eMBB-URLLC co-existence using various ML techniques has been provided in Table 8.

a: LESSONS LEARNED

Machine Learning (ML) aided resource allocation techniques have demonstrated better performance in the co-existence of eMBB and URLLC traffic. Specially, prediction of URLLC arrival by using NN and DNN greatly facilitates to correctly allocate the URLLC resources at an appropriate instant such that both eMBB throughput and URLLC latency requirements can be achieved [68]. Appropriate selection of TTI is decided using Random Forest (RF) based algorithm called RF-ETDA [64]. NN and DNN cannot predict the arrival of URLLC traffic in the dynamic situations as the CSI varies continuously due to the movement of users. DRL based algorithms can enhance the performance of dynamic placement of URLLC traffic inside on-going eMBB transmission based on reward received by the agent. Both single-agent and multi-agent DRL based techniques have been implemented with different objectives. In [66], DRL facilitates solving a non-convex optimization problem by getting a sub-optimal solutions of the converted convex problems. Furthermore, resource allocations due to state change of users can be addressed by MDP based DRL [66] techniques. Also, as the state changes are continuous, DDPG based DRL [69] assists for maintaining a trade-off between the heterogeneous service requirements for eMBB and URLLC.

E. DISTRIBUTED AND FEDERATED LEARNING BASED RESOURCE ALLOCATION

Distributed machine learning can be used in 6G networks so that highly dynamic, heterogeneous and multidimensional systems are possible. These systems in 6G need more adaptive, flexible and intelligent techniques that bring a revolutionary leap of communications with the sureness of massive broadband, ultra-reliable and low latency [75]. Additionally, resource management and access control are decided on the basis of extracted meaningful information from the enormous amount of data generated from the networks. In [76], a distributed ML algorithm called Long-Short-Term Memory (LSTM) has been used to estimate the true state based on past observations. This algorithm then facilitates the eMBB spectrum access dynamically by adopting real-world setting in more efficient way, and thus the issues of computational complexity of the problems with large state space can be mitigated. This work can be extended to the eMBB and URLLC co-existence so that the improved average channel utility and average throughput obtained will be used to compensate the degradation of QoS of the eMBB service due to stealing of the spectrum by the URLLC users. In [77], a Neighbor-Agent Actor-Critic (NAAC) model has been used as a multi-agent DRL to tackle the instabilities in the environment. In this model, eMBB and URLLC devices share the spectrum resources based on the trained information from respective nodes with the decentralized implementation but in centralized manner.

In the distributed learning approach, both off-line and on-line ML based resource management have been studied in the literature. Off-line resource management seems impractical for eMBB and URLLC systems as it assumes perfect channel state condition within the considered wireless networks which is not possible in many real-life scenarios. On the other hand, on-line resource management for the co-existence of various service classes in heterogeneous wireless networks provides a realistic framework in the same context. Resource allocation problems in on-line ML based frameworks use the stochastic manner with the modeling of channel state as MDP. There are two approaches for on-line resource management: centralized and distributed. Due to the severe feedback overheads, centralized approach is not applicable in the case when there are large number of users. Besides, there is also limitation of the applied MDP as it suffers from significant computational complexity that makes it intractable to function. On the other hand, fully distributed on-line resource allocation does not require any information exchange among the IoT devices. However, in the dynamic environment where the users are moving, such distributed on-line resource management system is not always convergent and hence it is not always easy to realize. This is because resource management strategies assume that the global system state is available for each device. Therefore, the authors in [78] have proposed a mean-field multi-agent DRL algorithm with an objective of maximizing throughput of considered IoT system. Some other works [79], [80], [81] have also used ML based algorithms for the on-line distributed IoT systems to manage the resource allocations applied to maximize the sum rate for the eMBB users and lower the latency in URLLC service. In the eMBB and URLLC co-existence, a low latency and the faster decision making is guaranteed via the distributed learning as the edge servers do not require to send data to a centralized server.

Some other works have utilized the distributed machine learning frameworks to efficiently accommodate both eMBB and URLLC services. For example, authors in [82] have presented a resource allocation scheme for multi-access edge-computing using Radio Access Technology (RAT). In this work, the authors used multi-agent DRL to meet the QoS requirements of both URLLC and eMBB users. In the proposed Distributed DRL (DDRL), the learning efficiency has been increased with the participation of the multiple actors. Compared to a single actor, multiple actors perform better for exploration and exploitation process so that the latency minimization and the service rate slicing of the URLLC users are improved without decreasing the Offloading Success Rate (OSR) of the eMBB users.

The authors in [83] have proposed a UE-MEC integrated resource allocation problem for eMBB and URLLC co-existence system using a DRL based distributed framework with an objective of maximizing System Utility (SU) in terms of Spectral Efficiency (SE) and Offloading Success Rate (OSR). Some existing works in distributed frameworks such as [84], [85] have considered a heavy workload situation, but these works could not solve the issue of queuing

TABLE 8. Research on ML based co-existence models for eMBB and URLLC services.

ML Techniques	Ref.	Objective	Results	Remarks
RF-ETDA	[64]	To accomplish TTI selection for each service.	URLLC delay is improved by 45.64%	RF-ETDA provides 95.93% URLLC accuracy and 98.93% eMBB accuracy
DRL	[65]	To maximize eMBB data rate subject to the URLLC reliability constraints.	eMBB reliability of higher than 90%	Convergence and accuracy are used to measure computation complexity
MDP based DRL	[66]	To improve eMBB throughput and to reduce URLLC loss rate	URLLC loss rate is reduced by 43.7%, 28.6% and 53.8% than random, greedy and sequence strategies respectively.	Short TTI size and flexibility in the frequency domain are balanced only with 60 kHz-0.25 ms and 30 kHz-0.5 ms systems.
DRL	[67]	To maximize the service level agreement (SLA) satisfaction degree of various types of users	An increased data rate is achieved with the proposed algorithm compared with MA-DQN algorithm.	Resource efficiency can be enhanced by introducing user intents into slicing resource allocation in case of sufficient resource available.
NN	[68]	To dynamically manage network slicing on per-user basis.	Provides 95% prediction accuracy with blocking and delay reduction by 50% and 30% respectively.	Performance is valid only for the high URLLC loads.
DRL-DDPG	[69]	To optimize overlapping URLLC positions and bandwidth allocations.	Reliability maximization of eMBB and URLLC by degrading individual QoS of minimum number of users.	Achieves the good trade-off between eMBB and the URLLC service requirement without degrading eMBB QoS.
Bayesian	[70]	To normalize eMBB throughput and to maximize URLLC utility.	Bayesian technique outperforms the random search and DRL technique.	URLLC latency requirement is not directly impacted by the complexity of the proposed Bayesian technique.
DRL	[71]	To dynamically allocate the incoming URLLC traffic by puncturing eMBB codewords using proximal policy optimization (PPO).	Code-words without an inner erasure code are included by using PPO agent thus proving versatility of the RL approach.	Both value-based and the policy-based algorithms have been used.
DL-NS	[73]	To maximize system throughput while meeting QoS requirements.	System throughput improved by 19% and 27% over regression-tree based slicing technique and equal allocation technique respectively at maximum delay of 1 ms.	Low complexity with convergence time is only 1730 s at 20 dB SNR level.
DRL	[74]	To balance KPIs in different heterogeneous services including URLLC and eMBB.	There is 29% improvement in power allocations and 21 times improvement in throughput.	Proposed algorithm experiences less than 0.5 ms latency degradation at 10^{-4} percentile.

delay for the URLLC users. Some other works [86], [87], [88] considered the optimization of bandwidth allocation and user resources to process different tasks such as URLLC and eMBB QoS management. According to [88], co-operation among multi-RATs is beneficial for Large Latency Sensitive Computing service tasks (L2SC) offloading of eMBB users with large task and minimization of the sum of the latency cost for the URLLC users. A local computing frequency in distributed manner with task splitting and transmit power are optimized with the formulated L2SC problem.

In the distributed resource allocation scheme, a global machine model is trained by using the network parameters at the edges and the global model shares the learning parameters at the network edge for distributive executions. A DRL based distributed algorithm can enhance the QoS of the eMBB users by ensuring the required URLLC reliability. A distributed learning approach in a multi-cell Fog-RAN architecture has been provided in [10] where the URLLC traffic is processed at the edge and the eMBB traffic is processed at the cloud. The authors utilized the Heterogeneous-NOMA (H-NOMA) to improve the latency for the URLLC service and spectral efficiency for the eMBB traffic. In the proposed distributed learning approach, redundant models can be trained and deployed across the various edge devices. In the eMBB and URLLC co-existence, the network performance is still alive with the minimum distraction even in the failure of some communication links thanks to the use of redundant models.

Federated Learning (FL) is considered as a co-operative learning approach in which ML models are trained with the help of multiple collaborators. These multiple collaborators constitute different wireless networks including Mobile Networks (MNOs), and use their private data-sets for training the models. A robust resource allocation model based on ML algorithm is generated when the global model is formed from aggregation entity, and there is further training opportunity for each collaborator after merging to the global model. Due to the heterogeneous requirement of QoS of various service classes such as eMBB and URLLC, radio resource allocation has become a complicated operation in mobile virtual networks. Radio Access Networks (RANs) slicing has been considered as a promising technique to overcome the resource allocation issues in heterogeneous networks requiring different QoS of various services.

Authors in [89] have proposed a RAN slicing mechanism based on DRL algorithm with an objective of allocating radio resources to the respective eMBB and URLLC users. To satisfy the QoS requirements of various types of users, the proposed mechanism requests for additional resources from other base stations when the existing base station is overloaded. The authors in [90] have proposed resource allocation scheme in a Multi-mobile Networks Operators (mNOs) cases where various networks need to satisfy the demands for different QoS. The authors in this work have used two level networks slicing based on a DRL-based framework to allocate resources to each network based on its own QoS requirements by maximizing user satisfaction rate.

Use of ML can significantly improve the training accuracy and energy efficiency of the wireless systems with the enormous number of IoT edge devices in the federated learning based wireless networks. As a result, communication latency of such edge networks will be improved [91]. The improved latency of the URLLC service will be maintained for the addition of the extra devices without degrading the QoS of the eMBB users because of the scalability of the federated learning based algorithms.

For the guarantee of the URLLC service, the authors in [92] have proposed an accelerated gradient-descent multiple access algorithm. The proposed algorithm has contributed to optimize the model accuracy and training speed. Another work in [93] has proposed a learning control algorithm to minimize the redundancy in the local data sets and CSI so that the local updates and the global parameters aggregation are matched resulting a reliable low latency communication.

A resource allocation problem in combination with the minimization of the training loss of federated learning has been presented in [94]. An appropriate user selection and resource allocation techniques used in this work have ensured the low latency communication. Another work in [95] has proposed an algorithm aiming to enhance the training speed, learning updates and the convergence. In the co-existence scenario, resource efficiency can be enhanced by training the data locally at the edges devices and thus reduce the latency of the URLLC users as there is no need to transfer the data to the cloud server. The cloud server is responsible for the processing of only eMBB services so that the system complexity will also be reduced.

The authors in [96] have investigated CSI in a low -latency federated learning framework whereas a device scheduling and resource allocation problem has been studied in [97] with the formulation of both channel conditions and local model updates to ensure the efficient low latency communication with federated learning approach. In the distributed and federated learning based resource allocation, changes in the network conditions are adapted by the learning algorithms. Network adaptability of the adopted distributed algorithm is crucial to ensure the QoS of the individual service class in the eMBB and URLLC co-existence because the CSI of the eMBB and URLLC users are subject to change due to device mobility and variations in the densities of the devices in the network scenario. A summary of the works using distributed and the federated learning based resource allocation techniques for eMBB and URLLC co-existence has been provided in Table 9.

a: LESSONS LEARNED

Resource allocations in eMBB and URLLC co-existence scenario will be highly benefited by the application of distributed and federated learning based frameworks. In particular, for the networks with large number of users and devices, distributed ML frameworks are useful to optimize the network performance with low latency and higher throughput. In this context, various works using distributed learning frameworks have been discussed in [75], [76], [77], [78], [79], [80], [81], [82], [83], [84], and [85], most of which have utilized appropriate ML techniques to meet the eMBB and URLLC service requirements. For the eMBB and URLLC co-existence scenario with the distributed learning, one of the two traffic models is trained at the edges and the other is trained at the centralized server so that network performance and QoS of each types of users can be improved thanks to the low complexity of the distributed algorithm. The computational complexity problem in large scale data set are solved by designing distributed learning algorithm. These algorithms are also more efficient, robust and scalable. Consequently, the QoS requirements of the heterogeneous network scenarios such as eMBB and URLLC services are easily fulfilled by using distributed learning algorithms. The works in [76], [77], and [78] have demonstrated the superiority of the distributed learning approaches over the centralized approach by significantly decreasing the latency and improving the data rates of various class of users. Federated learning approach also follows the same procedures as in the distributed learning approach. However, in the federated learning, each participant is able to initialize the training independently in the local level. In both federated and distributed learning approach, the local updates are sent to the cloud server that aggregates the global model after getting the information from the local levels. The advantage of these processes result for faster convergences than in the centralized approaches [89], [93], [96].

F. RIS-ASSISTED RESOURCE ALLOCATION

Performances of eMBB and URLLC services are direct consequence of channel quality for eMBB and URLLC users. Less resources are required for simultaneous transmission of eMBB and URLLC services in the favorable channel conditions. Less number of frequency resources are required to achieve the latency and reliability requirements for the

TABLE 9. Research on distributed and federated learning based resource allocation for eMBB and URLLC systems.

Ref.	Objective	Algorithm	Contributions
[76]	To maximize network utility in a distributed manner by using multi-user strategy for accessing the spectrum.	DRL based distributed dynamic spectrum access algorithm.	Proposed algorithm has enhanced learning polices in an online manner to each user in the absence of message exchange among users dealing with large state space.
[77]	To improve the reward performance and convergence of Cognitive Radio (CR) networks.	Multi-agent DRL based Neighbor-Agent Actor-Critic (NAAC) algorithm.	The proposed algorithm improves the convergence speed than Q-learning with ϵ-greedy algorithm with large networks.
[78]	To develop power control policy in on-line manner for Energy Harvesting (EH) networks in the presence of causal information about wireless channel.	Discrete-time Meanfield Game (MFG).	Both on-line and off-line learning policies have been adopted to establish the convergence by learning stationary solution.
[82]	To maximize System Utility (SU) and to minimize the energy consumption of UEs by improving spectral efficiency (SE) and Offloading Success Rate (OSR).	(RAT)-based partial offloading and Multi-access Edge Computing (MEC) algorithm.	The proposed scheme improves the latency and learning efficiency.
[83]	To minimize energy consumption and maximize throughput of eMBB user by maintaining URLLC QoS.	Block Co-ordinate Descent (BCD) based algorithm.	The proposed algorithms have increased the energy-efficiency of eMBB and URLLC users network by jointly adopting communication and computation resource allocation strategies.
[89]	To optimize the performance of URLLC and eMBB services by allocating radio resources to gNodeBs based on eMBB and URLLC QoS requirements.	Two time-scales RAN slicing algorithm.	The proposed algorithm improves the performance by improving the RB allocation and meeting the requirements of URLLC and eMBB QoS.
[93]	To minimize a loss function with the limited resource blocks for maintaining a trade-off between local update and global aggregation.	Gradient-descent based control algorithm.	Proposed scheme accelerates the convergence by managing the bounds with non-i.i.d. data distributions for federated learning.
[96]	To improve the Broadband Analog aggregation (BAA) on learning performance.	Federated edge learning (FEEL) algorithm.	There is a latency reduction with the proposed BAA technique as this outperforms the traditional OFMDA scheme in large users scenario.

punctured eMBB users in case of high channel gain. As a result, loss of data rate of the eMBB users will be also less. Similarly, in case of low channel gain, more frequency resources are required to achieve their target data rate. This results in less number of URLLC users to be served as the frequency resources are punctured to accommodate the URLLC traffic. Since improving channel gain is crucial for simultaneous eMBB and URLLC services, RIS is considered as one of the best technologies to improve the channel gain for solving the proposed problem.

RIS is considered as a promising technology for controlling and configuring the wireless propagation environment [98]. One of the major advantages of RIS is that it can amplify the received signal strength at the desired point by appropriately tuning the phase-shift of each reflecting passive element. Therefore, reflected signals by the RIS can be constructively added at the points of interests [99], [100]. As a result, the received signals as well as the channel gains at the eMBB and URLLC users are enhanced. Additionally, in context of co-existence of eMBB and URLLC service transmission, RIS can be a suitable technology due to the following reasons:

1) It can amplify received signals by passively reflecting the incident signals without exploiting extra frequency chains thus resulting lower power consumption.
2) It can create a LOS path for the signals propagation in the dead region where the direct signals are obstructed due to obstacles in dense urban area.
3) It can improve the channel gains by creating constructive interference of the reflected signals at the specific points where the eMBB and URLLC users are located in the disaster area.
4) Combination of RIS with UAV is able to achieve panoramic full-range reflection by adjusting the phase shift of RIS meta-elements which significantly increases the number of supported eMBB and URLLC mobile users.

As the wireless propagation environment can be controlled and configured with the help of using RIS, a big challenge arises on configuring the optimal phase shift of the RIS elements. Many works have been conducted in the literature on RIS phase shift optimization. For example, works in [101] and [102] have studied the RIS phase shift optimization. In [102], RIS-aided resource allocation is considered where the Orthogonal Frequency Division Multiple Access (OFDMA) technique is used to provide the resource allocations. In that work, a RIS-assisted system is proposed where a robust and secure resource allocation algorithm is implemented. However, works in [102] and [103] have considered only eMBB traffic for the resource allocation.

RIS-assisted URLLC system has been also investigated in some recent studies. URLLC sum rate maximization problem is solved in RIS-assisted OFMDA multi-cell system in [104]. Above mentioned works are related to either only eMBB or URLLC system. However, in [20], authors have solved the eMBB and URLLC co-existence problem using single antenna BS. The proposed optimization problem in [112] optimizes the RIS phase shift in each mini-slot thus causing extra delay for the URLLC users due to signalling overhead between RIS and BS. Additionally, optimizing phase shift of RIS meta-elements individually may cause higher complexity and high signalling overhead for large size RIS. The authors in [105] have utilized the optimized off-line code-book in an on-line stage so that signalling overhead depends on the RIS code-book size rather than on the RIS size or number of RIS meta-elements. Other works in [106] and [107] also provide RIS assisted wireless systems where the code-book based on-line optimization problem has been solved.

Use of RIS for eMBB and the URLLC co-existence is well presented in [107] where a code-book based optimization framework is adopted for the resource allocation optimization problem which is designed for the maximization of average throughput of eMBB users over time slots with satisfying URLLC QoS in each mini-slot. A sub-optimal solution was obtained by using an Alternating Optimization (AO) in an iterative fashion.

Almekhlafii et al. in [20] have used a RIS-assisted wireless communication system with the objective to enhance the system level performance of both types of services namely eMBB and URLLC in terms of URLLC reliability and eMBB throughout. The authors in that work provide eMBB and URLLC resource allocation strategies to achieve the overall objectives. The authors have proposed the solutions for both eMBB and URLLC resource allocations separately. As the eMBB resource allocation problem, a joint optimization of power allocation together with RIS phase shift is formulated to maximize the eMBB sum rate. Likewise, objective of the URLLC allocation problem is to maximize the URLLC admission packets and minimizing eMBB rate loss which is achieved by joint optimizing RIS phase shift, power allocations and frequency allocations. Since the URLLC delay is a critical aspect for the eMBB-URLLC co-existence, a novel framework needs is proposed to satisfy the URLLC latency requirements. To achieve this goal, URLLC reliability is enhanced by RIS phase shift matrix which is designed at the beginning of each slot. Two algorithms namely heuristic algorithm and the URLLC allocation algorithm are implemented for this purpose. It has been shown from the simulation that the proposed algorithms provide a low time complexity. The benefit of using RIS is justified by achieving 99.99% URLLC packets admission rate over 95.6% without using RIS. Also, there is also 70% improvement on the eMBB throughput.

In other work in [108], the authors have proposed RIS-aided eMBB and URLLC co-existence system where both eMBB and URLLC allocation schemes are studied. A double-stage AO algorithm is proposed for the eMBB allocation problem. On the other hand, problem of maximization of URLLC packets reception rate with the minimization of the eMBB rate loss are solved by satisfying the QoS of each class of service users. To pre-configure the RIS phase shift matrix, URLLC allocation algorithm utilizes the heuristic algorithm. Simulation results clearly demonstrate that nearly 95.5%. URLLC packet reception rate is achieved while maintaining the stringent URLLC latency requirement. On the other hand, there is only 6%. total eMBB rate loss due to URLLC service provision. In [109], Zarini et al. have studied a RIS-aided wide band terahertz (THz) communication system to meet eMBB and URLLC QoS requirements. A resource management problem is formulated with the purpose of jointly optimizing RIS phase shift, BS transmit power and THz resource block allocation. A supervised learning approach is proposed with the use of DL and ensemble learning methods. 49% improvement of spectral efficiency is achieved for the eMBB service with the use of large-sized RIS with satisfying latency and throughput requirements of each class of users. Moreover, a real-time resource management is achieved by using ensemble method at the expense of 1% performance loss. Table 10 presents brief outlines on RIS-assisted eMBB-URLLC co-existence system.

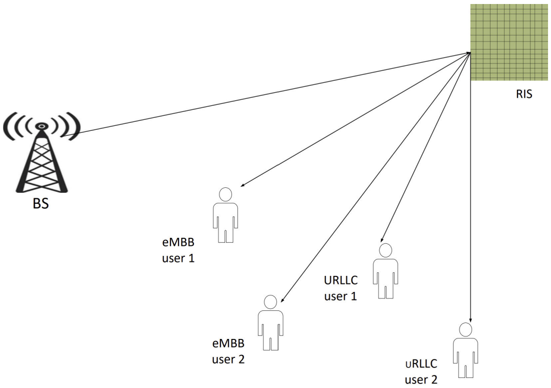


FIGURE 7. RIS-assisted eMBB and URLLC co-existence scenario.

1) CHALLENGES FOR ENABLING URLLC USING RIS

As we discussed above, use of RIS provides wide range of benefits such as energy efficiency and low cost by controlling wireless propagation environment. In [110], the authors have provided the potential advantages of deploying RIS in the upcoming 6G wireless networks for the co-existence of eMBB and URLLC services. In the co-existence model, it is important to meet the URLLC QoS requirements without impacting the eMBB performance. The authors, in that work, provide the relevant research directions of using RIS in the eMBB-URLLC system and have also pointed some challenges in RIS-aided eMBB-URLLC system. The authors have shown, through the simulation, that as the number of elements increases in RIS, URLLC reliability and eMBB data rate are also improved for various BS power allocation. This is because URLLC channel gain is improved for increased number of RIS elements. As a result, less number of frequency resources are needed to achieve the URLLC reliability and latency requirement without decreasing pre-allocated frequency resources to the eMBB users. However, following two challenges are pointed out by the authors when using RIS in co-existence of eMBB and URLLC services [111].

- Latency components of RIS.
- RIS deployment and BS-RIS-user association.

Latency and reliability requirements of URLLC traffic and eMBB throughput requirements are impacted by the additional processing delay due to estimating CSI and RIS phase shift optimization. Moreover, RIS deployment offers many challenges for meeting the relation between the RIS size, number of RIS meta-elements and associated QoS of both class of services [111]. Another challenge may arise as the different performance metrics such as coverage, latency, reliability, data rate and connectivity are to be guaranteed by using association strategies in large-sized-multi-RIS system.

2) POSSIBLE RESEARCH DIRECTIONS ON RIS ASSISTED EMBB AND URLLC SYSTEM

As discussed the associated challenges of using RIS with the eMBB and the URLLC system, the authors in [110] have suggested following research directions for the optimum utilization of the RIS.

- RIS densification: Deployments of large number of RISs within the same wireless networks architecture can be beneficial in the presence of high blockage density and serving many connected devices. This leverages flexibility in the RISs deployment and improves the LOS connectivity between BS and devices. Therefore, high reliability and lower outage probability is ensured for both eMBB and URLLC users.
- RIS-assisted multiple access schemes: Optimizing and configuring RIS for a proper scheduling for both types of users by improving the performances of both services is an important research direction. Adopting multiplexing techniques such as puncturing or superposition [22], [23], [24], [25], [26] can impact the RIS configuration and eMBB and URLLC scheduling.
- Leveraging ML tools: Prediction of the arrival of URLLC packets using ML techniques such as DML and FL can be an attractive solution for the case of RIS-aided co-existence of eMBB and URLLC services. Forecasting of such URLLC packets arrival facilitates the BS to correctly schedule the URLLC packets by jointly optimizing RIS phase shift and the power allocations so that eMBB rate loss can be minimized. Besides, this helps for estimating the CSI for URLLC users beforehand so that proactively optimization of RIS phase shift matrix is possible. This optimization ultimately alleviates the CSI estimation for ensuring URLLC latency requirement. %endenumerate

a: LESSONS LEARNED

RIS can be one of the important technologies for assisting eMBB and URLLC co-existence scenario. It can amplify the received signals at the point of interest so that total SNR and the throughput of the eMBB users can be improved. However, extra processing delay occurs for URLLC users due to optimizing the RIS phase shift and configuring CSI of the associated URLLC links. Therefore, there exists a strong trade-off between eMBB throughput and URLLC latency requirements. Aforementioned works on RIS assisted eMBB and URLLC system provided in [20], [103], [106], [108], and [109] have argued that eMBB service transmission is benefited thanks to the use of RIS rather than URLLC service transmission. However, these works have not addressed the URLLC latency requirements in detail. In order to leverage the full benefits of the RIS for URLLC users, appropriate multiplexing schemes need to be implemented with large number of RISs. Moreover, ML algorithms greatly enhance the benefits of RIS for URLLC traffic allocation by predicting the upcoming URLLC packets in mini slots thus ensuring the latency requirements of the URLLC traffic [110].

G. RIS AND UAV ASSISTED RESOURCE ALLOCATION

Among three important features (mMTC, eMBB and URLLC), URLLC is one of the most important and widely applicable features to be offered by upcoming 6G networks. Many mission critical applications such as V2V communications, UAVs control applications, autonomous driving, E-health, intelligent transportation, industrial applications and many others are supported by URLLC services [16]. Ultrahigh reliability with packet decoding error rate of 10^{-9} or less is required for such mission control applications. Again, for lowering the latency, such service class resorts to short packets. But, channel coding gain will be adversely effected by such short packets. Since the URLLC operates under finite block length regime, Shannon's capacity bound will be inapplicable in this case. In [112], channel coding rate and the finite block length have been studied for the provision of URLLC service.

UAVs, due to their flexible configuration, have been widely used for extending the wireless coverage in the specific scenario where the users are far away from the base station. Use of UAVs are more crucial when propagation of the signals are blocked. As a result, transmission efficiency will be improved in UAV assisted wireless communications. Since, RIS has been considered as an emerging technology for supporting high data rate with the minimum cost and less energy consumption [113], it can be combined with UAV for better results for various 6G applications.

URLLC has been considered as one of the major novel applications, and it is widely used in various fields such as environment monitoring, disaster communications, industrial automation and healthcare monitoring [114], [115]. In all such applications, milliseconds latency and 99.999% reliability are needed that can be only possible by the use of short

TABLE 10. Research on RIS assisted eMBB and URLLC Co-existence scenario.

Ref.	Service Type	Objective	Optimized parameters
[20]	eMBB and URLLC	URLLC admission packets maximization and eMBB rate loss minimization.	BS Power allocation and RIS phase-shift matrix.
[102]	eMBB	To maximize minimum rate of users.	RIS reflection coefficients, OFDMA time-frequency resource block and power allocations.
[103]	eMBB	To maximize secrecy rate of legitimate communication link.	Active transmit and passive reflect beamforming vectors.
[104]	URLLC	To guarantee the QoS of URLLC users and to maximize weighted system sum throughput.	Resource allocation and passive beamforming.
[105]	eMBB	To analyze fundamental trade-off between power efficiency and the size of the codebook.	RIS phase shift optimization.
[106]	eMBB and URLLC	To maximize average sum rate of eMBB users while ensuring QoS of each URLLC user.	Resource allocation and code-book optimization.
[107]	eMBB and URLLC	To enhance URLLC packets admission rate and eMBB average throughput.	RIS phase shift, power allocation and frequency allocations.
[108]	eMBB and URLLC	To meet the requirements of eMBB and URLLC QoS.	RIS phase shift, BS transmit power and THz resource block allocation.

packets transmission for certain QoS [116]. There are many research that have been progressed in RIS-UAV-URRLC applications. In [117], a decoding error minimization problem has been proposed by jointly optimizing UAV's location and the block-length allocation.

URLLC applications have gained a lot of flexibility and reliability by utilizing the RIS-UAV networks. In [118], the authors have proposed a RIS-UAV network for a multiple service provision requirements where the multiple services are provided concurrently instead of only one service at a time. The authors in [119] have presented a NOMA based cognitive UAV-URLLCs system that analyzes the provision of mMTC type of communication. RIS-UAV-URLLC system has been considered as a new paradigm for future 6G wireless communication system as this system copes with the significant increase in demand of the stringent QoS services. The authors in [120] have provided optimization framework for optimizing RIS passive beamforming, UAV position and block-length. Objective of this work is to minimize the decoding error rate. However, position of RIS is fixed as it has been deployed on a building. In [121], the authors have developed a joint optimization of RIS deployment and UAV trajectory for providing statistical delay and error rate bounded QoS schemes using FBC. In this work, FBC based effective energy maximization problem has been solved by using iterative algorithms.

In [118], [119], and [26], use of RIS has not been used. On the other hand, in other works [120], [121], and [122], the authors have used the RIS to meet the stringent QoS requirement of the URLLC system. However, in these works, stationary RIS has been used for providing URLLC services. The shortcoming of the stationary RIS is that the blocking effect still exists in such wireless networks. Aerial RIS can overcome such issue where the RIS is mounted on the UAV to significantly improve the LOS coverage as the signals can be provided from relatively higher altitude with the 3D mobility of the UAVs. Therefore, unlike the above mentioned RIS-UAV-URRLC systems, the authors in [123] have proposed the aerial RIS-URLLC systems where the RIS carrying UAVs move to improve the reliability and QoS of URLLC services in the area where the users are located far away from the base station. A zero-forcing beamforming and Time Division Multiple Access (TDMA) have been proposed to eliminate existing interference in the networks. The authors have optimized the user scheduling, UAV deployment, power allocations at BS, block length of URLLC and RIS phase shift. A Deep Neural Network (DNN) is used to solve the optimal UAV deployment followed by the resource allocation for maximization of reliability of the users with QoS constraints. Advantage of using the aerial RIS is to provide URLLC to massive number of IoT devices that are far away from the BS with an extension of the wireless coverage.

The authors in [121] have proposed a FBC based joint optimization framework for optimizing UAV trajectory and RIS deployment. Objective of the proposed work is to support the statistical delay and error rate bounded QoS for mURLLC traffic. With developing 3D channel models for mURLLC in FB regime, the authors have formulated an effective energy-efficiency maximization problem and solved it using an iterative algorithms using DRL. Research on RIS-UAV-URLLC system have been presented in Table 11.

a: LESSONS LEARNED

In the aforementioned works in UAV-URLLC system, the authors focus on URLLC using UAV with RIS or without RIS. Use of UAV can provide more flexibility for providing

LOS wireless signals, and thus contributes to improve the channel gain. Use of UAV can be a promising approach for the need of low cost and efficient real-time service demand fulfillment. They are appropriate for fast deployment, QoS requirement, monetary cost and user mobility [124]. UAV has an ability of facilitating on-demand deployment with cheap infrastructures to provide the faster connectivity support. Additionally, they have high flexibility of changing the positions to provide on-demand communications to ground users with LOS links [125], [126]. Combination of UAV with RIS can be an attractive solution for providing both eMBB and URLLC services. RIS-UAV system can alleviate eMBB throughput so that even if some of the frequency resources are stolen by the URLLC traffic, eMBB rate loss can be improved. However, mobility and trajectory management of UAVs are essential for providing uninterrupted and reliable wireless networks [127]. Moreover, challenges for deploying the RIS and optimizing RIS phase shift add more complexity for the functionality of URLLC system. In order to meet the stringent latency requirement of the URLLC traffic, appropriate ML techniques need to be leveraged to overcome the challenges associated with UAVs and RISs.

IV. APPLICATIONS ON CO-EXISTENCE OF EMBB AND URLLC IN 6G

As extreme low latency in milliseconds (ms) and high throughput in gigabits per second (Gpbs) are the features of 6G communication systems, many low latency and high throughput based applications are being emerged day by day. One of the widely used areas of eMBB and URLLC is the industrial internet of things (IIoT) where various mission critical applications are performed. Some of these applications include factory automation, health care systems, V2V and V2X communications, AR,VR and HD video, robotics, machine control system and many more. Each of these applications requires task specific QoS and throughput. Different communication technologies are used to enable these applications. Some of the important application areas of eMBB and URLLC system are discussed below.

A. FACTORY AUTOMATION

Factory automation involves the digitized works in factory system that help to eliminate the repetitive works in the machine environment and collect the operational data simultaneously. It is a basis of smart factory system where various mission critical parameters for closed loop control design demand extreme low latency of less than 1 ms with 99.999% reliability. Many URLLC enabled operations such as industrial robots are delay-sensitive and require extremely high service reliability. There are some important use cases in smart factory system that includes remote control monitoring, machine milling, machine for packaging, motion control and videoing. Among them, motion control and robotics are URLLC use cases [128] while Virtual Reality and Augmented Reality (AR and VR) are the eMBB use cases [129], [130]. All of these use cases require latency from 1 ms to 10 ms with reliability of 10^{-9} and up to 100 Mbps data rate.

B. HEALTH CARE SYSTEM

Improved health care system comprises of new technologies operated with IoT devices that have extra-ordinary sensing, ubiquitous identification and reliable communication abilities. All activities, objects and information are to be continuously tracked, monitored and updated in smart health care system. Many critical information related to medicine, equipment, diagnosis, therapy, logistics, treatment and activities of patients are to be processed, managed and shared with other system, authority and connections. In all of these tasks, very strict URLLC with minimum latency and high QoS and reliability are required. On the other hand, concept of IoT-based At-home Healthcare (AAH) [131] is now broadly accepted and applied to the places where the universal access of internet is guaranteed. Furthermore, for safety and security of the concerned activities, people, procedures and treatment, a reliable, seem-less and accessible communication services are to be provided. In addition, robot-assisted treatment and surgery not only demands extremely low end-to-end latency but also many remotely monitored patients and treatments should be fast, reliable and safe. These applications are possible only with the provision of URLLC and eMBB transmission.

C. AUTONOMOUS VEHICLE AND TRAFFIC CONTROL

Autonomous vehicles are operated with the assistance of many IoT sensors that collect, process and share very sensitive information. These information can be applied to control the motion of the vehicles. In order to ensure efficient and real-time exchange of large volume of data acquisition and processing, eMBB and URLLC enabled communication system is necessary with less than 5 ms latency and up to Mbps data rate. Furthermore, as self-driving vehicles shift the exchanged data to edge and cloud, these processes need to be done with high reliability. Various enabling technologies such as WiFi, ZigBee and B5G are used for secure and reliable data transfer in V2V and V2X communications. There are some important use cases in smart autonomous vehicle such as high speed, lane change and accidental warnings which seek quick data exchange and time sensitive information [132], [133], [134]. For efficient traffic control on high-ways and smart cities, low latency and high reliable communication enhance smart traffic management system. Effective traffic control is critical because it ensures the smooth movement across the cities and minimizes hazards and accidents [135].

D. SMART GRID

Thanks to 6G, functions of smart grid applications for electricity generation, transmission and distribution become more reliable, safe and scalable. For smart monitoring and controlling mechanism, IoT drivers in 6G contribute for additional reliability and low latency functionality. There

TABLE 11. Research on RIS assisted UAV-URLLC System.

Ref.	Objective	Use of RIS	RIS type
[26]	To minimize the completion time and energy consumption of UAV-URLLC system.	No	Unknown
[118]	To improve EE for eMBB coverage and URLLC QoS.	No	unknown
[119]	To maximize EE, minimize URLLC latency and enhance mMTC throughput.	No	unknown
[120]	To minimize total decoding error rate and find UAV's optimal position and blocklength.	Yes	Stationary RIS
[121]	To support statistical delay and error-rate bounded QoS for mURLLC.	Yes	Stationary RIS
[122]	To analyze the effect of imperfections in the positioning information and reliability.	Yes	Stationary RIS
[123]	To minimize the decoding error probability for URLLC users.	Yes	Mobile RIS

are various use cases in smart grid such as electricity distribution with different proportion that have different latency and throughput requirement as per the target needs. For the medium electricity distribution, a latency of 25 ms and reliability of 10^{-3} is required whereas high electricity distribution seeks for a latency of 5 ms and reliability of 10^{-6}. In both cases, up to 10 Mbps throughput is required. From consumer point of view, a latency of less than 5 ms and throughput up to 30 Mbps for uplink and 3 Mbps for downlink receptions are needed for any unexpected disturbance in usual grid stations. As a result, less than 1 ms latency and 99.99% reliability is ensured when 100 customers are simultaneously supported by smart grid. Use of 6G with mmwave predicts energy usage of customers more accurately and it transmits the information to micro-grid with higher reliability [132]. Moreover, there should be real-time power transmission from the grid that should be controlled and monitored with high precision. This operation can be achieved with the help of eMBB and URLLC service.

E. MINING INDUSTRY

Mining industry is one of the controlled and monitored applications of eMBB and URLLC. There should be continuous surveillance for the mining environment and worker's condition inside the mine. Use of IoT sensor is to collect data regarding the positions of mining workers each second so that safety and security of the workers are ensured. Various hazards and accidents such as rock sliding and suffocation can be correctly monitored and predicted using IoT-collected sensors data. In order to perform this operation, URLLC of stringent QoS with ms latency and high reliability is required. Furthermore, for the proper management of the mine monitoring system, eMBB traffic eases to collect the enormous mine related data at each instant [136].

F. ROBOTIC MOTION CONTROL AND NAVIGATION

Use of robots is increasing in various fields such as industry, health care system, inventory tracking and mining. In smart robotic system, collaboration of various robotic applications are enhanced with the help of eMBB and URLLC system where resilient and robust communication is guaranteed. Different wireless protocols including 5G/LTE, 802.11, 802.15.4 are used in various communication channels. Appropriate ML techniques can be implemented to improve the data analytic either for central processing or edge processing system so that safety and mission-critical robotic services can be improved. A detailed study on the robotic applications is provided in [137], [138], [139], and [140]. According to these works, URLLC latency of 4 to 8 ms with reliability of 10^{-9} and eMBB throughput up to 100 *Mbps* are needed to meet the requirements in the robust industrial robotic applications.

G. AUGMENTED REALITY (AR) AND VIRTUAL REALITY (VR)

VR is defined as an interactive simulation created by the computer software that provides an opportunity for a user to be engaged in an environment like the real world scenario. AR is a part of VR where the computer generated information is collaborated and interpreted on different objects, images, places to enhance user's learning experience. There are many applications areas where AR and VR are used to improve the learners ability to learn using digitized system. AR and VR are widely used in the health care education, school education and in the engineering innovation. The authors in [141], [142], [143], [144], and [145] have provided a detailed study on joint computing resource allocations and transmission for AR and VR. However, use of AR and VR demand very stringent QoS to map the virtual world on the real one which can be achieved by providing extremely high data rate traffic.

V. RESEARCH CHALLENGES ON EMBB-URLLC CO-EXISTENCE

Arrival of URLLC traffic is generally spontaneous and requires to be addressed quickly due to their latency constraints. The concept of resource reservation for these class of service is one of the solutions. However, this may cause to arise an issue of under-utilization of radio resources which demand highly effective multiplexing scheme. Commonly used multiplexing techniques that have been studied in literature are superposition/ puncturing [22], [23], [24], [25], [26] and flexible short TTI /puncturing [58], [59], [60], [61], [62]. Short TTI mechanism is beneficial in terms of less complexity but there is large amount of degradation of the spectral efficiency due to massive overhead for channel control. Puncturing mechanism can decrease the overhead in control channel but it requires a heavy attention to formulate the mechanism for compensating lost resource due to puncturing. To address the time sensitive transmission of the URLLC traffic, the concept of slots and mini slots are introduced where URLLC traffic are scheduled in mini slots on pre-scheduled eMBB traffic which degrades the QoS of eMBB users. In spite of paying much attention for satisfying the QoS of both class of users by sharing the spectrum between LTE and Wi-Fi [146], [147] and between LTE-A and NB-IoT [148], there is not much progress on fulfilling stringent QoS on URLLC while providing the sufficient throughput for the eMBB users. These works are limited on the system architecture and framework structure. Furthermore, most of the works in co-existence of eMBB and URLLC have focused on framework or system level design and analysis. Strong and concrete mathematical models for the co-existence of eMBB and URLLC traffic design are still missing. Again, the proposed mathematical models have high complexity relative to their contribution for fulfilling QoS requirements of both class of users. Therefore, development of appropriate mathematical model with minimum complexity and better QoS is a challenge for eMBB and URLLC co-existence system design.

One of the major challenges of eMBB and URLLC co-existence is to minimize the effect of URLLC transmission on the eMBB throughput due to resource sharing. It is almost impossible to nullify the eMBB throughput loss but it can be minimized by implementing suitable resource allocation schemes. Several optimization techniques have been implemented in the literature for scheduling URLLC traffic. However, optimization on URLLC transmission has always an impact on the eMBB user's QoS. Therefore, a joint optimization of eMBB and URLLC is highly recommended.

The impact of URLLC transmission on eMBB QoS can be minimized by improving the channel gain. However, there is always a question: how can we improve the channel condition and propagation environment? To answer this question, the authors in [20] have implemented RIS to enhance the channel gain of wireless link. But, there are several issues with RIS in terms of URLLC requirements because there is always an extra delay in the transmission due to RIS phase shift optimization. Use of RIS can enhance the eMBB throughput by improving channel gain, but there is additional complexity and delay due to RIS meta elements' phase angle design. Again, RIS power consumption model has not been derived yet in the literature. RIS phase shift optimization can hamper both energy efficiency and latency improvement. Similarly, use of UAV can also improve the channel gain by providing LOS link but finding optimal position of the UAV for providing LOS link adds more complexity in the system design. Furthermore, as UAV works on small battery power system, design of energy efficient UAV assisted wireless networks add extra complexity due to formulation of UAV position optimization problem.

Use of appropriate ML techniques with or without RIS-UAV assisted eMBB and URLLC co-existence system is an attractive solution. Specially, ML techniques are beneficial for predicting URLLC traffic arrival, and to schedule it in the mini slots. However, choice of appropriate ML technique is also a challenging task in dynamic network scenario where network parameters are constantly changed. Most of the ML based techniques have utilized the NN and DL [68] models but these models cannot correctly represent the fast changing network situations and, their results are not universally valid. Use of DRL [65] - [66], [68], [71], [74] based technique can improve the validity of the results but they possess additional complexity. Therefore, design of eMBB and URLLC co-existence system satisfying the QoS requirements of both eMBB and URLLC services is an open challenge in 6G wireless networks. The design considerations should include very efficient multiplexing schemes for improving the channel gain so that degradation of eMBB throughput can be minimized while satisfying QoS requirements of both class of users. Appropriate use of ML techniques with suitable wireless technologies are highly encouraged.

VI. SUMMARY

In this paper, we have provided a comprehensive survey on resource allocation schemes in eMBB and URLLC co-existence system in 6G. In particular, we have reviewed several research papers that have worked on the co-existence of eMBB and URLLC traffic using different resource allocation strategies such as network slicing, scheduling, flexible TTI, distributed and federated learning and machine learning based multiplexing techniques. Additionally, we also surveyed the research papers that have used RISs and UAVs on the co-existence system. Furthermore, we have explored the issues with eMBB and URLLC co-existence with RIS-UAV system and provided the possible solutions to these issues. Finally, we have pointed out some challenges and possible solution directions on the design of eMBB-URLLC co-existence systems.

REFERENCES

[1] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," *IEEE Netw.*, vol. 34, no. 3, pp. 134–142, May 2020.

[2] *Study on Physical Layer Enhancements for NR Ultra-Reliable and Low Latency Case (URLLC)*, 3GPP, document TR38.824, Mar. 2019.

[3] M. Almekhlafi, M. A. Arfaoui, C. Assi, and A. Ghrayeb, "Joint resource and power allocation for URLLC-eMBB traffics multiplexing in 6G wireless networks," in *Proc. ICC*, Oct. 2021, pp. 1–6, doi: 10.1109/ICC42927.2021.9500443.

[4] M. Alsenwi, S. R. Pandey, Y. K. Tun, K. T. Kim, and C. S. Hong, "A chance constrained based formulation for dynamic multiplexing of eMBB-URLLC traffics in 5G new radio," in *Proc. Int. Conf. Inf. Netw. (ICOIN)*, Kuala Lumpur, Malaysia, Jan. 2019, pp. 108–113.

[5] *Physical Channels and Signals for 5G-NR*, 3GPP, document TS 38.211, Aug. 2018.

[6] *Downlink Multiplexing of eMBB and uRLLC Transmission*, 3GPP, document R1-1700374, TSG RAN WG1 NR Ad-Hoc Meeting, Jan. 2017.

[7] M. Al-Ali and E. Yaacoub, ''Resource allocation scheme for eMBB and uRLLC coexistence in 6G networks,'' *Wireless Netw.*, vol. 29, no. 6, pp. 2519–2538, Aug. 2023.

[8] H. H. H. Mahmoud, A. A. Amer, and T. Ismail, ''6G: A comprehensive survey on technologies, applications, challenges, and research problems,'' *Eur. Trans. Telecommun.*, vol. 32, no. 4, pp. 1–14, 2021.

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

[10] R. Kassab, O. Simeone, P. Popovski, and T. Islam, ''Non-orthogonal multiplexing of ultra-reliable and broadband services in fog-radio architectures,'' *IEEE Access*, vol. 7, pp. 13035–13049, 2019.

[11] C. Xiao, J. Zeng, W. Ni, X. Su, R. P. Liu, T. Lv, and J. Wang, ''Downlink MIMO-NOMA for ultra-reliable low-latency communications,'' *IEEE J. Sel. Areas Commun.*, vol. 37, no. 4, pp. 780–794, Apr. 2019.

[12] E. J. dos Santos, R. D. Souza, J. L. Rebelatto, and H. Alves, ''Network slicing for URLLC and eMBB with max-matching diversity channel allocation,'' *IEEE Commun. Lett.*, vol. 24, no. 3, pp. 658–661, Mar. 2020.

[13] A. Dogra, R. K. Jha, and S. Jain, ''A survey on beyond 5G network with the advent of 6G: Architecture and emerging technologies,'' *IEEE Access*, vol. 9, pp. 67512–67547, 2021.

[14] Md. E. Haque, F. Tariq, M. R. A. Khandaker, K.-K. Wong, and Y. Zhang, ''A survey of scheduling in 5G URLLC and outlook for emerging 6G systems,'' *IEEE Access*, vol. 11, pp. 34372–34396, 2023.

[15] B. S. Khan, S. Jangsher, A. Ahmed, and A. Al-Dweik, ''URLLC and eMBB in 5G industrial IoT: A survey,'' *IEEE Open J. Commun. Soc.*, vol. 3, pp. 1134–1163, 2022.

[16] M. Bennis, M. Debbah, and H. V. Poor, ''Ultrareliable and low-latency wireless communication: Tail, risk, and scale,'' *Proc. IEEE*, vol. 106, no. 10, pp. 1834–1853, Oct. 2018.

[17] M. A. Siddiqi, H. Yu, and J. Joung, ''5G ultra-reliable low-latency communication implementation challenges and operational issues with IoT devices,'' *Electronics*, vol. 8, no. 9, p. 981, Sep. 2019.

[18] H. Ji, S. Park, J. Yeo, Y. Kim, J. Lee, and B. Shim, ''Ultra-reliable and low-latency communications in 5G downlink: Physical layer aspects,'' *IEEE Wireless Commun.*, vol. 25, no. 3, pp. 124–130, JUN. 2018.

[19] A. Anand, G. de Veciana, and S. Shakkottai, ''Joint scheduling of URLLC and eMBB traffic in 5G wireless networks,'' *IEEE/ACM Trans. Netw.*, vol. 28, no. 2, pp. 477–490, Apr. 2020.

[20] M. Almekhlafi, M. A. Arfaoui, M. Elhattab, C. Assi, and A. Ghrayeb, ''Joint resource allocation and phase shift optimization for RIS-aided eMBB/URLLC traffic multiplexing,'' *IEEE Trans. Commun.*, vol. 70, no. 2, pp. 1304–1319, Feb. 2022.

[21] F. Saggese, M. Moretti, and P. Popovski, ''Power minimization of downlink spectrum slicing for eMBB and URLLC users,'' *IEEE Trans. Wireless Commun.*, vol. 21, no. 12, pp. 11051–11065, Dec. 2022.

[22] 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.

[23] A. Karimi, K. I. Pedersen, N. H. Mahmood, G. Pocovi, and P. Mogensen, ''Efficient low complexity packet scheduling algorithm for mixed URLLC and eMBB traffic in 5G,'' in *Proc. IEEE 89th Veh. Technol. Conf.*, Apr. 2019, pp. 1–6.

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

[25] 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,'' 2020, *arXiv:2003.07651*.

[26] A. Manzoor, S. M. A. Kazmi, S. R. Pandey, and C. S. Hong, ''Contract-based scheduling of URLLC packets in incumbent EMBB traffic,'' 2020, *arXiv:2003.11176*.

[27] *TSG RAN WG1 Meeting 87*, document R1- 1612306, 3rd Generation Partnership Project (3GPP), Nov. 2016.

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

[29] K. I. Pedersen, M. Niparko, J. Steiner, J. Oszmianski, L. Mudolo, and S. R. Khosravirad, ''System level analysis of dynamic user-centric scheduling for a flexible 5G design,'' in *Proc. IEEE Global Commun. Conf.*, Dec. 2016, pp. 1–6.

[30] 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.

[31] A. Pradhan and S. Das, ''Joint preference metric for efficient resource allocation in co-existence of eMBB and URLLC,'' in *Proc. Int. Conf. Commun. Syst. Netw.*, 2020, pp. 897–899.

[32] X. Zhang, X. Guo, and H. Zhang, ''RB allocation scheme for eMBB and URLLC coexistence in 5G and beyond,'' *Wireless Commun. Mobile Comput.*, vol. 2021, pp. 1–7, Oct. 2021.

[33] Y. Prathyusha and T.-L. Sheu, ''Coordinated resource allocations for eMBB and URLLC in 5G communication networks,'' *IEEE Trans. Veh. Technol.*, vol. 71, no. 8, pp. 8717–8728, Aug. 2022.

[34] B. Shi, F.-C. Zheng, C. She, J. Luo, and A. G. Burr, ''Risk-resistant resource allocation for eMBB and URLLC coexistence under M/G/1 queueing model,'' *IEEE Trans. Veh. Technol.*, vol. 71, no. 6, pp. 6279–6290, Jun. 2022.

[35] A. Azari, M. Ozger, and C. Cavdar, ''Risk-aware resource allocation for URLLC: Challenges and strategies with machine learning,'' *IEEE Commun. Mag.*, vol. 57, no. 3, pp. 42–48, Mar. 2019.

[36] A. Zaki-Hindi, S.-E. Elayoubi, and T. Chahed, ''URLLC and eMBB coexistence in unlicensed spectrum: A preemptive approach,'' in *Proc. Int. Wireless Commun. Mobile Comput. (IWCMC)*, Jun. 2020, pp. 229–234.

[37] G. Mountaser, T. Mahmoodi, and O. Simeone, ''Reliable and low-latency fronthaul for tactile internet applications,'' *IEEE J. Sel. Areas Commun.*, vol. 36, no. 11, pp. 2455–2463, Nov. 2018.

[38] 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.

[39] Y. Zhao, X. Chi, L. Qian, Y. Zhu, and F. Hou, ''Resource allocation and slicing puncture in cellular networks with eMBB and URLLC terminals coexistence,'' *IEEE Internet Things J.*, vol. 9, no. 19, pp. 18431–18444, Oct. 2022.

[40] M. Castaneda, A. Mezghani, and J. A. Nossek, ''Optimal resource allocation in the downlink/uplink of singleuser MISO/SIMO FDD systems with limited feedback,'' in *Proc. IEEE 10th Workshop Signal Process. Adv. Wireless Commun.*, Feb. 2009, pp. 354–358.

[41] A. Matera, R. Kassab, O. Simeone, and U. Spagnolini, ''Non-orthogonal eMBB-URLLC radio access for cloud radio access networks with analog fronthauling,'' *Entropy*, vol. 20, no. 9, p. 661, Sep. 2018.

[42] P. Korrai, E. Lagunas, S. K. Sharma, S. Chatzinotas, A. Bandi, and B. Ottersten, ''A RAN resource slicing mechanism for multiplexing of eMBB and URLLC services in OFDMA based 5G wireless networks,'' *IEEE Access*, vol. 8, pp. 45674–45688, 2020.

[43] K. Zhang, X. Xu, J. Zhang, B. Zhang, X. Tao, and Y. Zhang, ''Dynamic multiconnectivity based joint scheduling of eMBB and uRLLC in 5G networks,'' *IEEE Syst. J.*, vol. 15, no. 1, pp. 1333–1343, Mar. 2021.

[44] D. Shen, T. Zhang, J. Wang, Q. Deng, S. Han, and X. S. Hu, ''QoS guaranteed resource allocation for coexisting eMBB and URLLC traffic in 5G industrial networks,'' in *Proc. IEEE 28th Int. Conf. Embedded Real-Time Comput. Syst. Appl. (RTCSA)*, Aug. 2022, pp. 81–90.

[45] G. Interdonato, S. Buzzi, C. D'Andrea, L. Venturino, C. D'Elia, and P. Vendittelli, ''On the coexistence of eMBB and URLLC in multi-cell massive MIMO,'' *IEEE Open J. Commun. Soc.*, vol. 4, pp. 1040–1059, 2023.

[46] Q. Chen, H. Jiang, and G. Yu, ''Service oriented resource management in spatial reuse-based C-V2X networks,'' *IEEE Wireless Commun. Lett.*, vol. 9, no. 1, pp. 91–94, Jan. 2020.

[47] E. J. D. Santos, R. D. Souza, and J. L. Rebelatto, ''Rate-splitting multiple access for URLLC uplink in physical layer network slicing with eMBB,'' *IEEE Access*, vol. 9, pp. 163178–163187, 2021.

[48] M. Almekhlafi, M. Chraiti, M. A. Arfaoui, C. Assi, A. Ghrayeb, and A. Alloum, ''A downlink puncturing scheme for simultaneous transmission of URLLC and eMBB traffic by exploiting data similarity,'' *IEEE Trans. Veh. Technol.*, vol. 70, no. 12, pp. 13087–13100, Dec. 2021.

[49] S. Bakri, P. A. Frangoudis, and A. Ksentini, "Dynamic slicing of RAN resources for heterogeneous coexisting 5G services," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2019, pp. 1–6.

[50] M. Setayesh, S. Bahrami, and V. W. S. Wong, "Joint PRB and power allocation for slicing eMBB and URLLC services in 5G C-RAN," in *Proc. IEEE Global Commun. Conf.*, Dec. 2020, pp. 1–6.

[51] Y. Liu, B. Clerckx, and P. Popovski, "Network slicing for eMBB, URLLC, and mMTC: An uplink rate-splitting multiple access approach," *IEEE Trans. Wireless Commun.*, vol. 14, no. 8, pp. 1–14, Aug. 2021.

[52] A. A. Esswie and K. I. Pedersen, "Opportunistic spatial preemptive scheduling for URLLC and eMBB coexistence in multi-user 5G networks," *IEEE Access*, vol. 6, pp. 38451–38463, 2018.

[53] A. K. Bairagi, Md. S. Munir, M. Alsenwi, N. H. Tran, S. S. Alshamrani, M. Masud, Z. Han, and C. S. Hong, "Coexistence mechanism between eMBB and uRLLC in 5G wireless networks," *IEEE Trans. Commun.*, vol. 69, no. 3, pp. 1736–1749, Mar. 2021.

[54] A. A. Esswie and K. I. Pedersen, "Capacity optimization of spatial preemptive scheduling for joint URLLC-eMBB traffic in 5G new radio," in *Proc. IEEE Globecom Workshops*, Dec. 2018, pp. 1–6.

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

[56] 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.

[57] C.-Y. Chen, G.-L. Hung, and H.-Y. Hsieh, "A study on a new type of DDoS attack against 5G ultra-reliable and low-latency communications," in *Proc. Eur. Conf. Netw. Commun. (EuCNC)*, Dubrovnik, Croatia, Jun. 2020, pp. 188–193, doi: 10.1109/EUCNC48522.2020.9200956.

[58] N. Kumar and A. Ahmed, "ANFIS-based reactive strategy for uRLLC and eMBB traffic multiplexing in 5G new radio," in *Advances in Communication and Computational Technology*. Singapore: Springer, 2019, pp. 1409–1419.

[59] G. Pocovi, K. I. Pedersen, B. Soret, M. Lauridsen, and P. Mogensen, "On the impact of multi-user traffic dynamics on low latency communications," in *Proc. Int. Symp. Wireless Commun. Syst. (ISWCS)*, Sep. 2016, pp. 204–208.

[60] 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)*, Dec. 2016, pp. 1–7.

[61] 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.

[62] S. Dutta, M. Mezzavilla, R. Ford, M. Zhang, S. Rangan, and M. Zorzi, "Frame structure design and analysis for millimeter wave cellular systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1508–1522, Mar. 2017.

[63] Z. Zhang, Y. Gao, Y. Liu, and Z. Li, "Performance evaluation of shortened transmission time interval in LTE networks," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2018, pp. 1–5.

[64] J. Zhang, X. Xu, K. Zhang, B. Zhang, X. Tao, and P. Zhang, "Machine learning based flexible transmission time interval scheduling for eMBB and uRLLC coexistence scenario," *IEEE Access*, vol. 7, pp. 65811–65820, 2019.

[65] 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.

[66] C. Tang, X. Chen, Y. Chen, and Z. Li, "Dynamic resource optimization based on flexible numerology and Markov decision process for heterogeneous services," in *Proc. IEEE 25th Int. Conf. Parallel Distrib. Syst. (ICPADS)*, Dec. 2019, pp. 610–617.

[67] M. A. Riad, O. El-Ghandour, and A. M. A. El-Haleem, "Joint user-slice pairing and association framework based on H-NOMA in RAN slicing," *Sensors*, vol. 22, no. 19, p. 7343, Sep. 2022.

[68] N. Kumar and A. Ahmad, "Machine learning-based QoS and traffic-aware prediction-assisted dynamic network slicing," *Int. J. Commun. Netw. Distrib. Syst.*, vol. 28, no. 1, p. 27, 2022.

[69] J. Li and X. Zhang, "Deep reinforcement learning-based joint scheduling of eMBB and URLLC in 5G networks," *IEEE Wireless Commun. Lett.*, vol. 9, no. 9, pp. 1543–1546, Sep. 2020.

[70] W. Zhang, M. Derakhshani, G. Zheng, C. S. Chen, and S. Lambotharan, "Bayesian optimization of queuing-based multichannel URLLC scheduling," *IEEE Trans. Wireless Commun.*, vol. 22, no. 3, pp. 1763–1778, Mar. 2023.

[71] P. He, L. Zhao, S. Zhou, and Z. Niu, "Water-filling: A geometric approach and its application to solve generalized radio resource allocation problems," *IEEE Trans. Wireless Commun.*, vol. 12, no. 7, pp. 3637–3647, Jul. 2013.

[72] F. Saggese, L. Pasqualini, M. Moretti, and A. Abrardo, "Deep reinforcement learning for URLLC data management on top of scheduled eMBB traffic," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2021, pp. 1–6.

[73] K. Suh, S. Kim, Y. Ahn, S. Kim, H. Ju, and B. Shim, "Deep reinforcement learning-based network slicing for beyond 5G," *IEEE Access*, vol. 10, pp. 7384–7395, 2022.

[74] M. Elsayed and M. Erol-Kantarci, "AI-enabled radio resource allocation in 5G for URLLC and eMBB users," in *Proc. IEEE 2nd 5G World Forum (5GWF)*, Sep. 2019, pp. 590–595.

[75] J. Du, C. Jiang, J. Wang, Y. Ren, and M. Debbah, "Machine learning for 6G wireless networks: Carrying forward enhanced bandwidth, massive access, and ultrareliable/low-latency service," *IEEE Veh. Technol. Mag.*, vol. 15, no. 4, pp. 122–134, Dec. 2020.

[76] 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, doi: 10.1109/TWC.2018.2879433.

[77] Z. Li and C. Guo, "Multi-agent deep reinforcement learning based spectrum allocation for D2D underlay communications," *IEEE Trans. Veh. Technol.*, vol. 69, no. 2, pp. 1828–1840, Feb. 2020, doi: 10.1109/TVT.2019.2961405.

[78] M. K. Sharma, A. Zappone, M. Assaad, M. Debbah, and S. Vassilaras, "Distributed power control for large energy harvesting networks: A multi-agent deep reinforcement learning approach," *IEEE Trans. Cognit. Commun. Netw.*, vol. 5, no. 4, pp. 1140–1154, Dec. 2019, doi: 10.1109/TCCN.2019.2949589.

[79] L. Xiao, X. Wan, X. Lu, Y. Zhang, and D. Wu, "IoT security techniques based on machine learning: How do IoT devices use AI to enhance security?" *IEEE Signal Process. Mag.*, vol. 35, no. 5, pp. 41–49, Sep. 2018, doi: 10.1109/MSP.2018.2825478.

[80] A. Ferdowsi and W. Saad, "Deep learning for signal authentication and security in massive internet-of-Things systems," *IEEE Trans. Commun.*, vol. 67, no. 2, pp. 1371–1387, Feb. 2019, doi: 10.1109/TCOMM.2018.2878025.

[81] A. Stamou, N. Dimitriou, K. Kontovasilis, and S. Papavassiliou, "Autonomic handover management for heterogeneous networks in a future internet context: A survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3274–3297, 4th Quart., 2019, doi: 10.1109/COMST.2019.2916188.

[82] J. Yun, Y. Goh, W. Yoo, and J.-M. Chung, "5G multi-RAT URLLC and eMBB dynamic task offloading with MEC resource allocation using distributed deep reinforcement learning," *IEEE Internet Things J.*, vol. 9, no. 20, pp. 20733–20749, Oct. 2022.

[83] Y. K. Tun, D. H. Kim, M. Alsenwi, N. H. Tran, Z. Han, and C. S. Hong, "Energy efficient communication and computation resource slicing for eMBB and URLLC coexistence in 5G and beyond," *IEEE Access*, vol. 8, pp. 136024–136035, 2020.

[84] F. Sufyan and A. Banerjee, "Computation offloading for distributed mobile edge computing network: A multiobjective approach," *IEEE Access*, vol. 8, pp. 149915–149930, 2020.

[85] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, Y.-C. Liang, and D. I. Kim, "Applications of deep reinforcement learning in communications and networking: A survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3133–3174, 4th Quart., 2019.

[86] C. She, Y. Duan, G. Zhao, T. Q. S. Quek, Y. Li, and B. Vucetic, "Cross-layer design for mission-critical IoT in mobile edge computing systems," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 9360–9374, Dec. 2019.

[87] R. Dong, C. She, W. Hardjawana, Y. Li, and B. Vucetic, "Deep learning for hybrid 5G services in mobile edge computing systems: Learn from a digital twin," *IEEE Trans. Wireless Commun.*, vol. 18, no. 10, pp. 4692–4707, Oct. 2019.

[88] M. Qin, N. Cheng, Z. Jing, T. Yang, W. Xu, Q. Yang, and R. R. Rao, "Service-oriented energy-latency tradeoff for IoT task partial offloading in MEC-enhanced multi-RAT networks," *IEEE Internet Things J.*, vol. 8, no. 3, pp. 1896–1907, Feb. 2021.

[89] A. Filali, Z. Mlika, S. Cherkaoui, and A. Kobbane, "Dynamic SDN-based radio access network slicing with deep reinforcement learning for URLLC and eMBB services," *IEEE Trans. Netw. Sci. Eng.*, vol. 9, no. 4, pp. 2174–2187, Jul. 2022.

[90] G. Chen, X. Zhang, F. Shen, and Q. Zeng, "Two tier slicing resource allocation algorithm based on deep reinforcement learning and joint bidding in wireless access networks," *Sensors*, vol. 22, no. 9, p. 3495, May 2022.

[91] J. Du, B. Jiang, C. Jiang, Y. Shi, and Z. Han, "Gradient and channel aware dynamic scheduling for over-the-air computation in federated edge learning systems," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 4, pp. 1035–1050, Apr. 2023.

[92] R. Paul, Y. Friedman, and K. Cohen, "Accelerated gradient descent learning over multiple access fading channels," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 2, pp. 532–547, Feb. 2022.

[93] S. Wang, T. Tuor, T. Salonidis, K. K. Leung, C. Makaya, T. He, and K. Chan, "Adaptive federated learning in resource constrained edge computing systems," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1205–1221, Jun. 2019.

[94] M. Chen, Z. Yang, W. Saad, C. Yin, H. V. Poor, and S. Cui, "A joint learning and communications framework for federated learning over wireless networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 269–283, Jan. 2021.

[95] R. Balakrishnan, T. Li, T. Zhou, N. Himayat, V. Smith, and J. Bilmes, "Diverse client selection for federated learning via sub-modular maximization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Mar. 2022, pp. 1–18.

[96] G. Zhu, Y. Wang, and K. Huang, "Broadband analog aggregation for low-latency federated edge learning," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 491–506, Jan. 2020.

[97] D. Liu, G. Zhu, J. Zhang, and K. Huang, "Data-importance aware user scheduling for communication-efficient edge machine learning," *IEEE Trans. Cognit. Commun. Netw.*, vol. 7, no. 1, pp. 265–278, Mar. 2021.

[98] M. Di Renzo, A. Zappone, M. Debbah, M.-S. Alouini, C. Yuen, J. de Rosny, and S. Tretyakov, "Smart radio environments empowered by reconfigurable intelligent surfaces: How it works, state of research, and the road ahead," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2450–2525, Nov. 2020.

[99] E. Basar, M. Di Renzo, J. De Rosny, M. Debbah, M. Alouini, and R. Zhang, "Wireless communications through reconfigurable intelligent surfaces," *IEEE Access*, vol. 7, pp. 116753–116773, 2019.

[100] M. Samir, M. Elhattab, C. Assi, S. Sharafeddine, and A. Ghrayeb, "Optimizing age of information through aerial reconfigurable intelligent surfaces: A deep reinforcement learning approach," *IEEE Trans. Veh. Technol.*, vol. 70, no. 4, pp. 3978–3983, Apr. 2021.

[101] Q. Wu and R. Zhang, "Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5394–5409, Nov. 2019.

[102] Y. Yang, S. Zhang, and R. Zhang, "IRS-enhanced OFDMA: Joint resource allocation and passive beamforming optimization," *IEEE Wireless Commun. Lett.*, vol. 9, no. 6, pp. 760–764, Jun. 2020.

[103] X. Yu, D. Xu, Y. Sun, D. W. K. Ng, and R. Schober, "Robust and secure wireless communications via intelligent reflecting surfaces," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2637–2652, Nov. 2020.

[104] W. R. Ghanem, V. Jamali, and R. Schober, "Joint beamforming and phase shift optimization for multicell IRS-aided OFDMA-URLLC systems," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Mar. 2021, pp. 1–7.

[105] M. Najafi, V. Jamali, R. Schober, and H. V. Poor, "Physics-based modeling and scalable optimization of large intelligent reflecting surfaces," *IEEE Trans. Commun.*, vol. 69, no. 4, pp. 2673–2691, Apr. 2021.

[106] V. Jamali, M. Najafi, R. Schober, and H. V. Poor, "Power efficiency, overhead, and complexity tradeoff of IRS codebook design—Quadratic phase-shift profile," *IEEE Commun. Lett.*, vol. 25, no. 6, pp. 2048–2052, Jun. 2021.

[107] W. R. Ghanem, V. Jamali, M. Schellmann, H. Cao, J. Eichinger, and R. Schober, "Codebook based two-time scale resource allocation design for IRS-assisted eMBB-URLLC systems," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2022, pp. 419–425.

[108] X. Shen, Z. Zeng, and X. Liu, "RIS-assisted network slicing resource optimization algorithm for coexistence of eMBB and URLLC," *Electronics*, vol. 11, no. 16, p. 2575, Aug. 2022.

[109] H. Zarini, N. Gholipoor, M. R. Mili, M. Rasti, H. Tabassum, and E. Hossain, "Resource management for multiplexing eMBB and URLLC services over RIS-aided THz communication," *IEEE Trans. Commun.*, vol. 71, no. 2, pp. 1207–1225, Feb. 2023.

[110] M. Almekhlafi, M. A. Arfaoui, C. Assi, and A. Ghrayeb, "Enabling URLLC applications through reconfigurable intelligent surfaces: Challenges and potential," *IEEE Internet Things Mag.*, vol. 5, no. 1, pp. 130–135, Mar. 2022, doi: 10.1109/IOTM.007.2100124.

[111] C. Pan, H. Ren, K. Wang, J. F. Kolb, M. Elkashlan, M. Chen, M. Di Renzo, Y. Hao, J. Wang, A. L. Swindlehurst, X. You, and L. Hanzo, "Reconfigurable intelligent surfaces for 6G systems: Principles, applications, and research directions," *IEEE Commun. Mag.*, vol. 59, no. 6, pp. 14–20, Jun. 2021.

[112] Y. Polyanskiy, H. V. Poor, and S. Verdu, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[113] C. Pan, H. Ren, K. Wang, M. Elkashlan, A. Nallanathan, J. Wang, and L. Hanzo, "Intelligent reflecting surface aided MIMO broadcasting for simultaneous wireless information and power transfer," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1719–1734, Aug. 2020.

[114] D. Van Huynh, S. R. Khosravirad, L. D. Nguyen, and T. Q. Duong, "Multiple relay robots-assisted URLLC for industrial automation with deep neural networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Madrid, Spain, Dec. 2021, pp. 1–5.

[115] D. V. Huynh, S. R. Khosravirad, L. D. Nguyen, K. K. Nguyen, and T. Q. Duong, "Industrial IoTs clustering, joint blocklength and power optimisation for relay robots-aided URLLC in factory automation," *IEEE Internet Things J.*, vol. 12, no. 8, pp. 1–5. Feb. 2022.

[116] G. J. Sutton, J. Zeng, R. P. Liu, W. Ni, D. N. Nguyen, B. A. Jayawickrama, X. Huang, M. Abolhasan, Z. Zhang, E. Dutkiewicz, and T. Lv, "Enabling technologies for ultra-reliable and low latency communications: From PHY and MAC layer perspectives," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 2488–2524, 3rd Quart., 2019.

[117] C. Pan, H. Ren, Y. Deng, M. Elkashlan, and A. Nallanathan, "Joint blocklength and location optimization for URLLC-enabled UAV relay systems," *IEEE Commun. Lett.*, vol. 23, no. 3, pp. 498–501, Mar. 2019.

[118] P. Yang, X. Xi, K. Guo, T. Q. S. Quek, J. Chen, and X. Cao, "Proactive UAV network slicing for URLLC and mobile broadband service multiplexing," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 10, pp. 3225–3244, Oct. 2021.

[119] S. R. Sabuj, A. Ahmed, Y. Cho, K.-J. Lee, and H.-S. Jo, "Cognitive UAV-aided URLLC and mMTC services: Analyzing energy efficiency and latency," *IEEE Access*, vol. 9, pp. 5011–5027, 2021.

[120] A. Ranjha and G. Kaddoum, "URLLC facilitated by mobile UAV relay and RIS: A joint design of passive beamforming, blocklength, and UAV positioning," *IEEE Internet Things J.*, vol. 8, no. 6, pp. 4618–4627, Mar. 2021.

[121] X. Zhang, J. Wang, and H. V. Poor, "Joint optimization of IRS and UAV-trajectory: For supporting statistical delay and error-rate bounded QoS over mURLLC-driven 6G mobile wireless networks using FBC," *IEEE Veh. Technol. Mag.*, vol. 17, no. 2, pp. 55–63, Jun. 2022, doi: 10.1109/MVT.2022.3158047.

[122] F. Saggese, F. Chiariotti, K. Kansanen, and P. Popovski, "Efficient URLLC with a reconfigurable intelligent surface and imperfect device tracking," in *Proc. IEEE Int. Conf. Commun.*, May 2023, pp. 1–9.

[123] Y. Li, C. Yin, T. Do-Duy, A. Masaracchia, and T. Q. Duong, "Aerial reconfigurable intelligent surface-enabled URLLC UAV systems," *IEEE Access*, vol. 9, pp. 140248–140257, 2021.

[124] M. Mozaffari, W. Saad, M. Bennis, Y.-H. Nam, and M. Debbah, "A tutorial on UAVs for wireless networks: Applications, challenges, and open problems," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 2334–2360, 3rd Quart., 2019.

[125] A. Merwaday and I. Guvenc, "UAV assisted heterogeneous networks for public safety communications," in *Proc. IEEE Wireless Commun. Netw. Conf. Workshops (WCNCW)*, Mar. 2015, pp. 329–334.

[126] D. Liu, Y. Xu, J. Wang, Y. Xu, A. Anpalagan, Q. Wu, H. Wang, and L. Shen, "Self-organizing relay selection in UAV communication networks: A matching game perspective," *IEEE Wireless Commun.*, vol. 26, no. 6, pp. 102–110, Dec. 2019.

[127] L. Ferranti, F. Cuomo, S. Colonnese, and T. Melodia, "Drone cellular networks: Enhancing the quality of experience of video streaming applications," *Ad Hoc Netw.*, vol. 78, pp. 1–12, Sep. 2018.

[128] J. Yang, B. Ai, I. You, M. Imran, L. Wang, K. Guan, D. He, Z. Zhong, and W. Keusgen, "Ultra-reliable communications for industrial Internet of Things: Design considerations and channel modeling," *IEEE Netw.*, vol. 33, no. 4, pp. 104–111, Jul. 2019.

[129] J. Cheng, W. Chen, F. Tao, and C.-L. Lin, "Industrial IoT in 5G environment towards smart manufacturing," *J. Ind. Inf. Integr.*, vol. 10, pp. 10–19, Jun. 2018.

[130] E. J. Khatib and R. Barco, "Optimization of 5G networks for smart logistics," *Energies*, vol. 14, no. 6, p. 1758, Mar. 2021.

[131] M. C. Domingo, "An overview of the Internet of Things for people with disabilities," *J. Netw. Comput. Appl.*, vol. 35, no. 2, pp. 584–596, Mar. 2012.

[132] S. M. A. A. Abir, A. Anwar, J. Choi, and A. S. M. Kayes, "IoT-enabled smart energy grid: Applications and challenges," *IEEE Access*, vol. 9, pp. 50961–50981, 2021.

[133] P. Duan, Y. Jia, L. Liang, J. Rodriguez, K. M. S. Huq, and G. Li, "Space-reserved cooperative caching in 5G heterogeneous networks for industrial IoT," *IEEE Trans. Ind. Informat.*, vol. 14, no. 6, pp. 2715–2724, Jun. 2018.

[134] Z. Li, M. A. Uusitalo, H. Shariatmadari, and B. Singh, "5G URLLC: Design challenges and system concepts," in *Proc. 15th Int. Symp. Wireless Commun. Syst. (ISWCS)*, Aug. 2018, pp. 1–6.

[135] P. Schulz, M. Matthe, H. Klessig, M. Simsek, G. Fettweis, J. Ansari, S. A. Ashraf, B. Almeroth, J. Voigt, I. Riedel, A. Puschmann, A. Mitschele-Thiel, M. Muller, T. Elste, and M. Windisch, "Latency critical IoT applications in 5G: Perspective on the design of radio interface and network architecture," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 70–78, Feb. 2017.

[136] A. Aziz, O. Schelén, and U. Bodin, "A study on industrial IoT for the mining industry: Synthesized architecture and open research directions," *IoT*, vol. 1, no. 2, pp. 529–550, Dec. 2020.

[137] C. Li, C.-P. Li, K. Hosseini, S. B. Lee, J. Jiang, W. Chen, G. Horn, T. Ji, J. E. Smee, and J. Li, "5G-based systems design for tactile internet," *Proc. IEEE*, vol. 107, no. 2, pp. 307–324, Feb. 2019.

[138] F. Voigtländer, A. Ramadan, J. Eichinger, C. Lenz, D. Pensky, and A. Knoll, "5G for robotics: Ultra-low latency control of distributed robotic systems," in *Proc. Int. Symp. Comput. Sci. Intell. Controls (ISCSIC)*, Oct. 2017, pp. 69–72.

[139] J. Ansari, I. Aktas, C. Brecher, C. Pallasch, N. Hoffmann, M. Obdenbusch, M. Serror, K. Wehrle, and J. Gross, "Demo: A realistic use-case for wireless industrial automation and control," in *Proc. Int. Conf. Networked Syst. (NetSys)*, Mar. 2017, pp. 1–2.

[140] A. Fellan, C. Schellenberger, M. Zimmermann, and H. D. Schotten, "Enabling communication technologies for automated unmanned vehicles in industry 4.0," in *Proc. Int. Conf. Inf. Commun. Technol. Converg. (ICTC)*, Oct. 2018, pp. 171–176.

[141] Y. Zhou, C. Pan, P. L. Yeoh, K. Wang, M. Elkashlan, B. Vucetic, and Y. Li, "Communication-and-computing latency minimization for UAV-enabled virtual reality delivery systems," *IEEE Trans. Commun.*, vol. 69, no. 3, pp. 1723–1735, Mar. 2021.

[142] L. Teng, G. Zhai, Y. Wu, X. Min, W. Zhang, Z. Ding, and C. Xiao, "QoE driven VR 360° video massive MIMO transmission," *IEEE Trans. Wireless Commun.*, vol. 21, no. 1, pp. 18–33, Jan. 2022.

[143] X. Wei, C. Yang, and S. Han, "Prediction, communication, and computing duration optimization for VR video streaming," *IEEE Trans. Commun.*, vol. 69, no. 3, pp. 1947–1959, Mar. 2021.

[144] J. Ren, Y. He, G. Huang, G. Yu, Y. Cai, and Z. Zhang, "An edge-computing based architecture for mobile augmented reality," *IEEE Netw.*, vol. 33, no. 4, pp. 162–169, Jul. 2019.

[145] G. S. Park, R. Kim, and H. Song, "Collaborative virtual 3D object modeling for mobile augmented reality streaming services over 5G networks," *IEEE Trans. Mobile Comput.*, early access, Feb. 8, 2022, doi: 10.1109/TMC.2022.3149543.

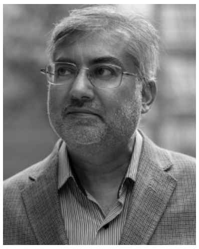
[146] A. K. Bairagi, S. F. Abedin, N. H. Tran, D. Niyato, and C. S. Hong, "QoE-enabled unlicensed spectrum sharing in 5G: A game-theoretic approach," *IEEE Access*, vol. 6, pp. 50538–50554, 2018.

[147] A. K. Bairagi, N. H. Tran, W. Saad, and C. S. Hong, "Bargaining game for effective coexistence between LTE-U and Wi-Fi systems," in *Proc. IEEE/IFIP Netw. Operations Manage. Symp.*, Apr. 2018, pp. 1–8.

[148] S. Liu, F. Yang, J. Song, and Z. Han, "Block sparse Bayesian learning-based NB-IoT interference elimination in LTE-advanced systems," *IEEE Trans. Commun.*, vol. 65, no. 10, pp. 4559–4571, Oct. 2017.

BHAGAWAT ADHIKARI (Member, IEEE) was born in Shantinagar, Jhapa, Nepal. He received the B.Sc. and M.Sc. degrees from Tribhuvan University, Nepal, in 2001 and 2004, respectively, and the B.Eng. degree in electrical and computer engineering and the M.A.Sc. degree in electrical engineering from Ryerson University, Toronto, Canada, in 2019. He is currently pursuing the Ph.D. degree in electrical engineering with Toronto Metropolitan University, Toronto, Canada.

He is an active member of the WINCORE Laboratory, Toronto Metropolitan University. His research interests include 6G wireless communication, indoor positioning, eMBB and URLLC services, signal processing, machine learning, and artificial intelligence. He is working on RIS and UAV assisted wireless communication systems as his research.



MUHAMMAD JASEEMUDDIN (Member, IEEE) received the B.E. degree from NED University, Pakistan, the M.S. degree from The University of Texas at Arlington, and the Ph.D. degree from the University of Toronto. He worked with the Advanced IP Group and the Wireless Technology Laboratory (WTL), Nortel Networks. He is currently a Professor and the Program Director of the Computer Networks Program, Toronto Metropolitan University. His research interests include network automation, caching in 5G and ICN networks, context-aware mobile middleware and mobile cloud, localization, power-aware MAC and routing for sensor networks, heterogeneous wireless networks, and IP routing and traffic engineering.

ALAGAN ANPALAGAN (Senior Member, IEEE) received the B.A.Sc., M.A.Sc., and Ph.D. degrees in electrical engineering from the University of Toronto, Canada.

He joined the ECBE Department, Toronto Metropolitan University (formerly Ryerson University), Canada, in 2001, where he was promoted to Full Professor in 2010. He was with the department in administrative positions as the Associate Chair, the Program Director of electrical engineering, and the Graduate Program Director. During his sabbatical, he was a Visiting Professor with the Asian Institute of Technology and a Visiting Researcher with Kyoto University. His industrial experiences include working for three years with Bell Mobility, Nortel Networks, and IBM. He directs a research group working on radio resource management (RRM) and radio access and networking (RAN) areas within the WINCORE Laboratory. He was a recipient of the IEEE Canada J. M. Ham Outstanding Engineering Educator Award, in 2018; the YSGS Outstanding Contribution to Graduate Education Award, in 2017; the Deans Teaching Award, in 2011; and the Faculty Scholastic, Research and Creativity Award thrice from Ryerson University. He is a Registered Professional Engineer in the province of Ontario, Canada, and Fellow of Institute of Electrical and Electronics Engineers (FIEEE). He served as the TPC Co-Chair for IEEE VTC Fall 2017, IEEE INFOCOM 2016, IEEE GLOBECOM 2015, and IEEE PIMRC 2011. He served as the Vice Chair for the IEEE SIG on Green and Sustainable Networking and Computing with Cognition and Cooperation, from 2015 to 2018; the IEEE Canada Central Area Chair, from 2012 to 2014; the IEEE Toronto Section Chair, from 2006 to 2007; the ComSoc Toronto Chapter Chair, from 2004 to 2005; and the IEEE Canada Professional Activities Committee Chair, from 2009 to 2011. He served as an Editor for IEEE COMMUNICATIONS SURVEYS AND TUTORIALS, from 2012 to 2014; IEEE COMMUNICATIONS LETTERS, from 2010 to 2013; and *EURASIP Journal of Wireless Communications and Networking*, from 2004 to 2009. He also served as the Guest Editor for six special issues published in IEEE, IET, and ACM.