Computer Communications xxx (xxxx) 1–28

Contents lists available at ScienceDirect

Computer Communications

journal homepage: www.elsevier.com

QoS aware resource allocation for coexistence mechanisms between eMBB and URLLC: Issues, challenges, and future directions in 5G

Rajesh Kumar [a], Deepak Sinwar [a, *], Vijander Singh [b, c]

[a] Department of Computer and Communication Engineering, Manipal University Jaipur, India
[b] Department of Computer Science and Engineering, Manipal University Jaipur, India
[c] Department of ICT and Science, Faculty of Information Technology and Electrical Engineering, NTNU – Norwegian University of Science and Technology, Ålesund, Norway

ARTICLE INFO

Keywords:
eMBB
URLLC
Quality of services
QoS
Scheduling
Coexistence
5G

ABSTRACT

5G NR enables three types of use case scenarios viz. enhance mobile broadband (eMBB), ultra-reliable low latency communication (URLLC), and massive machine type communication (mMTC). The eMBB is suitable for applications that demand higher throughput. Whereas URLLC is suitable for mission-critical applications with stringent requirement of low latency and reliability. The mMTC on the other hand is suitable for machine-to-machine (M2M) communications and massive IoT infrastructures. To meet all these requirements, 5G NR combines eMBB with URLLC services under a unified 5G air interface framework. Most of the coexistence mechanisms between eMBB and URLLC were presented by the researchers with a goal to enhance eMBB throughput with stringent latency and reliability requirements of URLLC services. Formulation of optimal resource scheduling and allocation were found to be the key problems of eMBB and URLLC traffic. This paper investigated the 5G state-of-the-art focused on coexistence mechanisms between eMBB and URLLC traffic for resource scheduling. We followed the PRISMA statement and classified the works to five major classes viz. multiplexing-, QoS-, Machine learning-, Network slicing-, and C-RAN architecture-based approaches. Each work was carefully examined against their methodology and performance metrics. In addition, several key issues, challenges, and future directions were also highlighted to provide detailed insights for researchers working in the field of 5G.

1. Introduction

5G mobile technology brings a revolutionary era in the next-generation mobile technology. After the rollout of the 5G services, several countries announced to prompt 5G applications in a wide range of industries, smart cities, smart healthcare, AR/VR, and to create a susceptible ecosystem for the growth of their countries [1]. 5G connectivity is the fuel of the economy for any country that promises to lead consumers, industries, and governments to new digital innovation for the productivity and growth of the economy [2]. It opens a new world of possibilities for every tech firm and provides intelligent automation and industry digitalization [3]. 5G utilizes millimeter wave technology for higher data rate transfer. As per 3GPP frequency bands, 5G occupies a range of frequencies in the spectrum that is broadly categorized into two frequency ranges viz. FR1 and FR2. The frequency band of FR1 remains below 7.125 GHz, whereas in FR2 it is above 24 GHz. FR1 is one of the traditional cellular mobile communications as compared to FR2. The FR2 utilizes millimeter wave technology for short-range and higher data rate transfer capabilities [4]. 5G supports three types of use case scenarios and several new services as compared to the 4G mobile technology viz., (1) enhanced mobile broadband (eMBB), (2) ultra-reliable low latency communication (URLLC), and (3) massive machine type of communication mMTC [5,6]. Some of the use cases require multiple dimensions for optimization, whereas some others are focused on key performance indicators [7,8]. eMBB services require a high data rate hence throughput is one of the key performance parameters. URLLC is focused on reliability and low latency, whereas mMTC deals with connecting billions of devices that support reliable data transfer capabilities [9]. To satisfy these use cases and services, 5G enables a diverse framework for self-adaptation, scalability, virtualization platform, reconfigurability, and self-organization network capabilities [10]. URLLC services are given the highest priority among other services, here user experience data rate is 25 Mbps and E2E latency is 1 ms. eMBB is provided with the second highest priority which supports a data rate of 100 Mbps and E2E latency of 10 ms. The third priority in this class is given to mMTC services where the user experienced data rate is of 100 Kbps with an E2E

* Corresponding author.
E-mail addresses: rajeshkaswa@gmail.com (R. Kumar), deepak.sinwar@gmail.com (D. Sinwar), vijander.singh@ntnu.no (V. Singh).

https://doi.org/10.1016/j.comcom.2023.10.024
Received 24 January 2023; Received in revised form 14 July 2023; Accepted 30 October 2023
0140-3664/

latency of less than 5 ms [11]. The coexistence of eMBB and URLLC is one of the challenging tasks because URLLC is not maintaining any queue for scheduling the traffic, so we can say it is sporadic, hence a dynamic multiplexing scheme is required for the coexistence of eMBB and URLLC services [12]. 3GPP proposed a dynamic framework called the puncturing/superposition technique for scheduling the coexistence traffic [13]. It utilizes the concept of preemption for scheduling URLLC traffic while serving eMBB traffic. There is another puncturing approach as mentioned in 3GPP called short TTI for dynamic multiplexing of eMBB and URLLC traffic in 5G system [14]. This puncturing technique has lower overhead but it requires efficient management and recovery of the puncturing slot. On the other hand, the short TTI technique faces high control channel overhead and low spectrum utilization. Based on superposition and puncturing techniques, researchers contribute a lot to enhance the capability of eMBB while maintaining URLLC latency and reliability constraints.

The coexistence mechanism between eMBB and URLLC has motivated us to accomplish this comprehensive review and to find out various issues and challenges that are big hurdle to implement end applications. We followed PRISMA statement [15] to prepare this review of eMBB and URLLC services. Research articles were fetched majorly from well-known publishers viz. IEEE, Elsevier, Springer, ACM, MDPI, etc. Key phrases for selecting relevant articles were, "eMBB and URLLC in the 5G network", "coexistence mechanism between eMBB and URLLC", "resource allocation in 5G″, etc. Around 200 sources related to the key phrase items were selected for preparing this review as shown in Fig. 1.

We examined the work embodied in these articles and framed several issues, challenges, and future direction related to the eMBB and URLLC coexistence mechanism. In short, the major contribution of this paper is as follows.

- In-depth review of various research papers to frame an extensive literature review based on QoS provisioning resource allocation between eMBB and URLLC coexistence mechanism.
- Presented 3GPP releases for 5G NR and related research concerning real-life deployments.
- Classified the coexistence mechanisms on various granularity levels viz. multiplexing, QoS, network slicing, machine learning, and C-RAN architecture.
- Presented various popular simulation parameters used in related works.
- Highlight various issues, open research challenges, and future directions for coexistence between eMBB and URLLC services.

The rest of the paper is organized as follows: Section 2 describes the 5G NR frame, multiple numerologies, SCS, dynamic multiplexing approach for eMBB and URLLC, QoS architecture, and resource scheduling approaches. In addition, section 2 also present 3GPP release for coexistence of eMBB and URLLC. Section 3 presents an extensive literature and discussions on the coexistence mechanism between eMBB and URLLC by classifying coexistence approaches into five main classes. Various simulation parameters used by the researchers for the coexis-

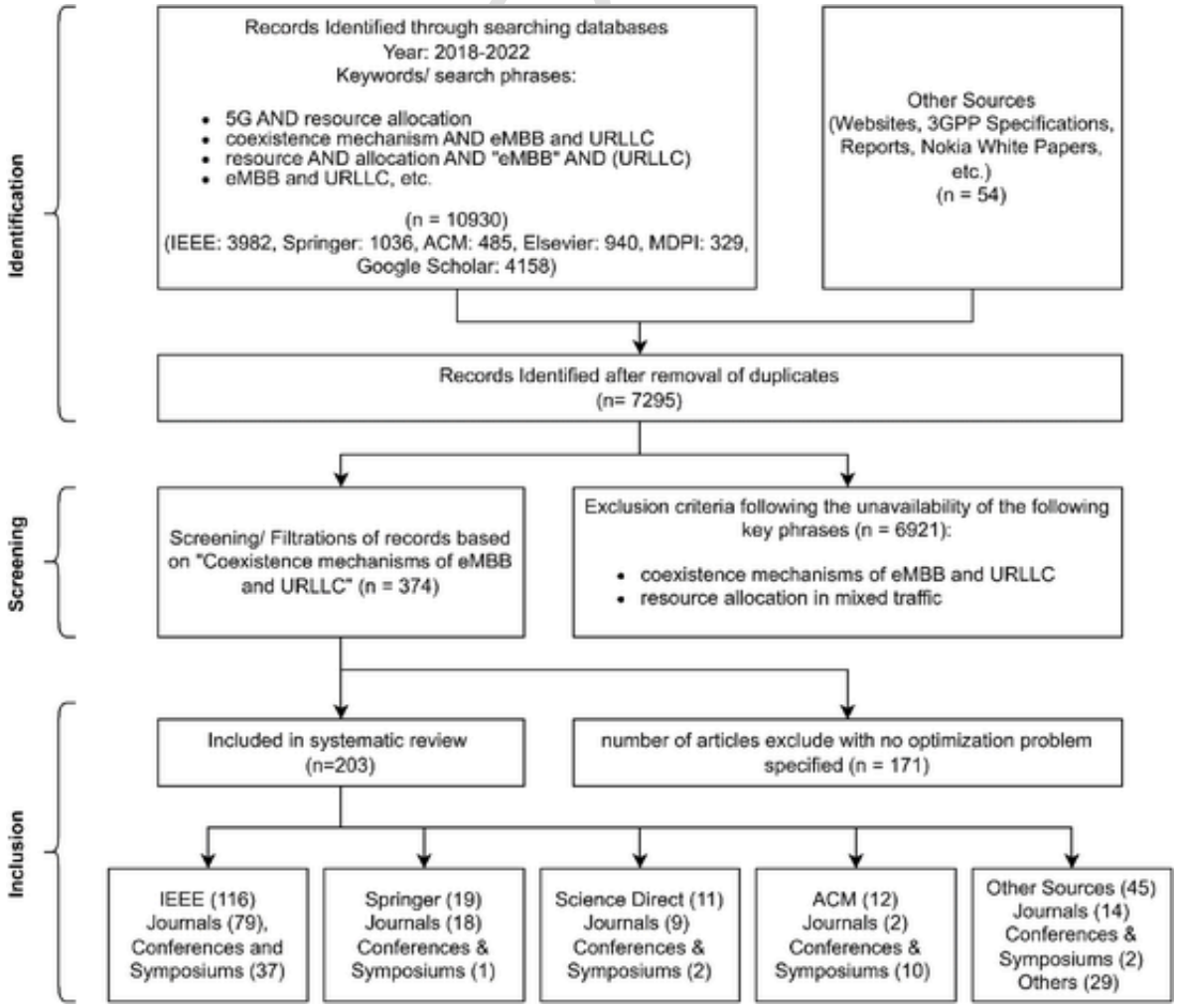

Fig. 1. Strategy followed for inclusion of sources for this study.

tence in the selected literature are highlighted in section 4. Section 5 presents a few use-case scenarios with reference to real life deployments. Section 6 highlights several key issues, open research challenges and future direction for coexistence mechanism between eMBB and URLLC services. Finally, section 7 concludes the survey.

2. 5G NR multiple numerology and subcarrier spacing

According to 3GPP specifications, Numerology (μ) is defined as subcarrier spacing types. In LTE, there was only a single subcarrier spacing that uses 15 kHz frequency band, but in 5G NR five different levels of μ (0–4) are used to denote the subcarrier frequency type (Δf) as shown in Table 1 [16,17]. 5G NR is utilizing different operating frequency bands such as 6 GHz, 10 GHz, and 28 GHz (millimeter waves); due to this, working with a single subcarrier spacing type is not suitable, hence flexible numerology is used in 5G NR [18]. 5G NR is supporting OFDMA for downlink transmission, here user's resources are dynamically multiplexed in both time and frequency domain grid. The frame structure of 5G NR is divided in such a way that the frame length is 10 ms for each frame and each subframe is defined as 1 ms as shown in Fig. 2 [19]. Here slot length differentiates based on numerology used by the user [20]. Each time slot contains a fixed 14 OFDMA symbol, the basic unit of resource element (RE) is consisting of 1 OFDMA symbol and 1 Subcarrier [21]. Thus, for $\mu = 0$, subcarrier is 12 kHz (all μ have same subcarrier value of 12 kHz) that occupy $12 \times 15 = 180$ kHz total space in

Table 1
5G NR multiple numerology and sub carrier spacing.

μ	$\Delta f = 2^{\mu}$. 15 KHz	Cyclic prefix	No of symbol in one slot	Slot Length	Subframe Length	No of Slot in one Subframe	Frame Length	Bandwidth per RB = 12*Δf
0	15	Normal	14	1 ms	1 ms	1	10	180 KHz
1	30	Normal	14	0.5 ms	1 ms	2	10	360 KHz
2	60	Normal/Extended	14/12	0.25 ms	1 ms	4	10	720 KHz
3	120	Normal	14	0.125 ms	1 ms	8	10	1440 KHz
4	240	Normal	14	0.625 ms	1 ms	16	10	2880 KHz

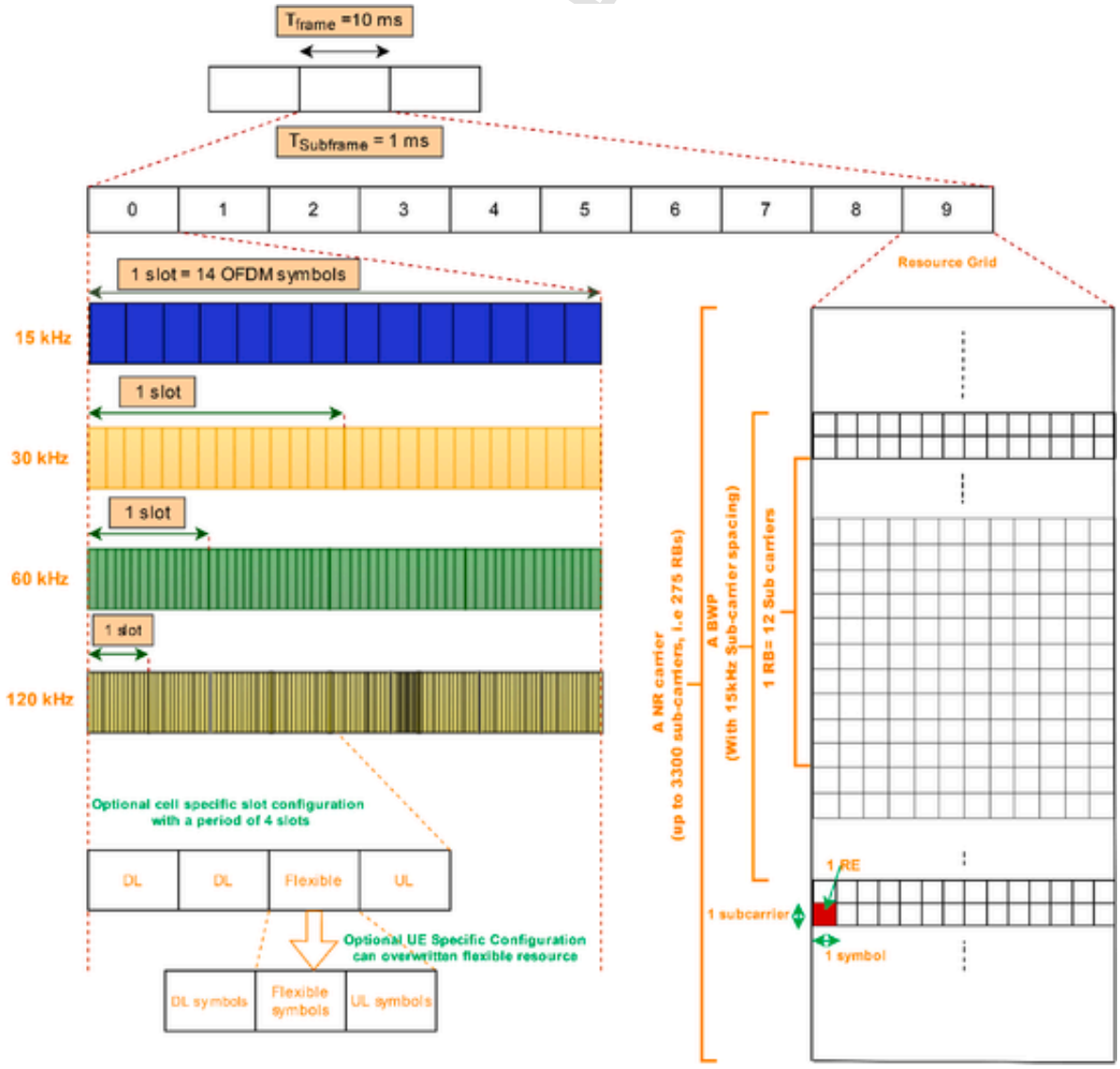

Fig. 2. Illustration 5G NR frame structure.

the frequency domain (denoted as PRB) and 12x14 = 168 total resource element (RE) in one physical resource block (PRB) [16].

2.1. Channel state information reporting method

For effective radio resource scheduling and allocation of the radio resources among UEs, an effective channel state information reporting method is required. Channel State Information (CSI) consists of Channel Quality Indicator (CQI) and precode matrix indicator, where CQI depends on receiving signal strength, interference, and noise ratio [22]. 5G NR supports three types of channel reporting methods in the physical uplink control channel (PUCCH). The first type of method is the wideband reporting method. In the second method, UEs select the subband with the best subband channel quality, and the third method adopts the higher layer configured subband reporting method. In the third method, gNB divides the downlink band into equal size subbands and UEs report the CQI value for each subband [23]. According to Ref. [24], the third method is more appropriate to achieve higher transmission efficiency.

2.2. DL multiplexing between URLLC and eMBB

This multiplexing technique is used to provide a co-existence region for both URLLC and eMBB traffic. gNB divides the Channel into frequency and time domains for effective allocation of the resources for eMBB and URLLC users. Here, eMBB users occupy the slot (1 Ms) for 15 kHz while scheduling the resource at beginning of the time slot, whereas, URLLC users are scheduled in the coexistence region of eMBB and URLLC as shown in Fig. 3 [25]. The base station (gNB) divides the channel into “the slot” and “mini slot” for occupying eMBB and URLLC user traffic. Here URLLC user traffic is urgent and needs to be immediately scheduled, which means there is no queue for URLLC users [25]. The numerology and frame structure of 5G provides services to meet such expectations and provide a coexistence region of URLLC and eMBB in the channel for scheduling the URLLC traffic.

2.3. Dynamic sharing of the resource using pre-emption techniques

Hybrid scheduling algorithm is involved in the dynamic sharing of the resource between eMBB and URLLC users. Here, two types of approaches are being followed for replacement of the slots.

- Pre-emption via slot puncturing
- Pre-emption with delayed transmission

2.3.1. Pre-emption via slot puncturing

A pre-emption approach schedules the resources based on the pre-empted resource allocation strategy. In case of pre-emption via puncture, the existing scheduling mechanism will be followed for eMBB users, which means the resource is scheduled as per the allocation schedule [26]. Whenever the URLLC traffic is introduced that has a higher precedence sequence and need to be schedule immediately, the eMBB slot will be punctured and URLLC traffic will be scheduled in the mini-slots. Puncturing mechanisms have a higher loss for eMBB users, but here schedule traffic is put in a pre-emption slot, so the loss for eMBB users is minimized as compared to without pre-emption scheduling [27].

2.3.2. Pre-emption with delayed transmission

In this approach, eMBB traffic will be halted whenever the URLLC traffic is inserted in the eMBB slot. But this halting mechanism is temporary, after finishing the URLLC traffic, the eMBB traffic is resumed. However, the eMBB data at the end of the originally scheduled data is not transmitted. eMBB users need to be notified when the data will be halted and resumed [28]. The pre-emption approach still faced degrading in the performance of the eMBB users because of slot puncturing. Some mechanisms will be adopted to recover the losses to some extent [29]. Automatic supplementary transmission mechanism is automatically scheduled by gNB whenever eMBB user read by an indication that the data byte was corrupted. The victim eMBB users find the supplementary transmission schedule by gNB. The victim eMBB users have already reallocated resources before the URLLC traffic that will be scheduled in the supplementary transmission block [30]. Fast scheduling capability is required for scheduling the resources in the supplementary block. The victim eMBB users are not able to decode the corrupted

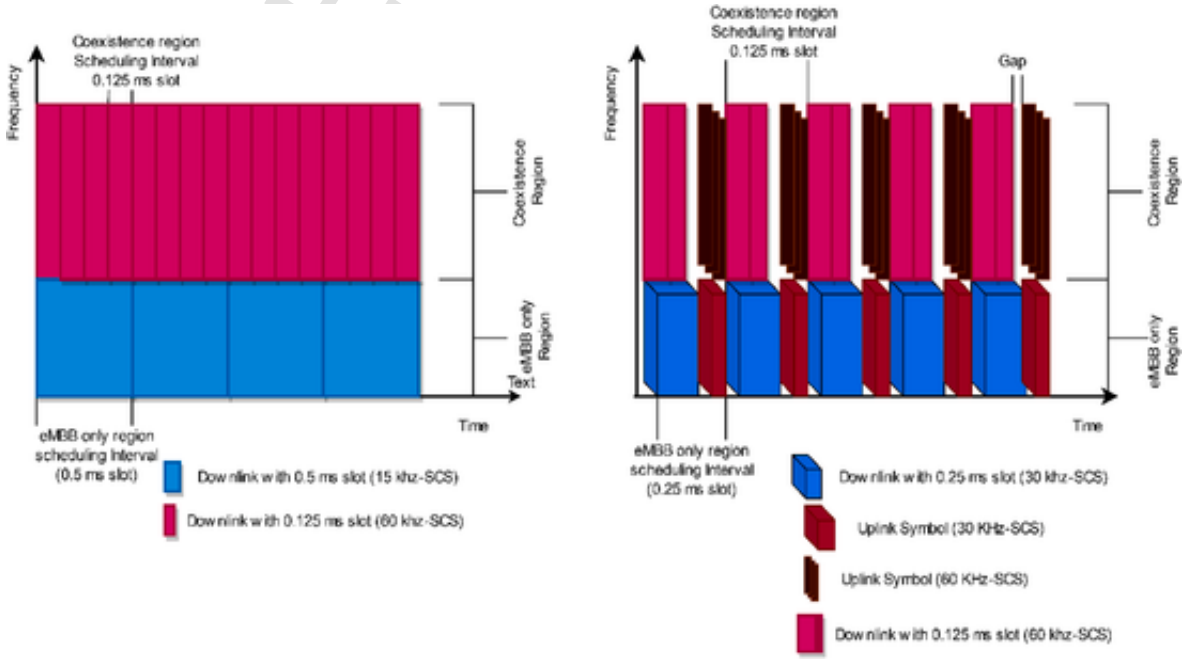

Fig. 3. FDM between eMBB only region and coexistence region.

transmission, hence it will reschedule in the supplementary transmission [31].

2.4. 5G QoS architecture

5G NR has different services requirements as compared to 4G network, hence several new features as well as parameters are introduced in the QoS architecture of 5G. QoS model is based on the unique QoS flow identifier, which represents the finest granularity to differentiate the level of QoS flow in the 5G system [32].

According to the packet treatment classification and packet detection rules, in a downlink direction whenever an IP data flow arrives at the UPF level from the data network, a unique QoS flow identifier (QFI) is being allocated, that uniquely identifies QoS flow in a particular PDU session as shown in Fig. 4 [33]. This process is known as the first level of mapping between the UPF and gNB [31].

According to the QoS marked and associated QoS profile, the second level of mapping is held between gNB and UEs at service data adaptation protocol (SDAP) layer, which lies between QoS flow and data radio barriers (DRBs) [34]. PDU Session is unique to UPF and UEs, the DRBs can hold multiple QoS flows in a particular PDU session. As per 3GPP specification, there are various types of QoS flows viz. QoS flow with GBR, QoS flow with NGBR, Delay Critical GBR, Reflective QoS (new for 5G). On the other hand, several QoS parameters are 5G QoS Identifier (5QI), allocation and retention priority, reflective QoS attributes (RQA), Notification Control, Flow bit rates, aggregates bit rates, Maximum packet loss rate, etc. [35]. In addition, there are several differences in the QoS flow architecture of 5G as compared to the 4G network. One of the major differences is that the 4G network have one to one mapping in QoS flow (UEs, eNB, and UPF), whereas 5G supports a two-step mapping: the first level mapping between UPF and gNB and the second level mapping between gNB and UEs. Other parameters are slightly differ from the 4G network as mentioned in Table 2 [36].

2.5. Radio resource scheduling

5G NR radio resources are scheduled by gNB to UEs once the DRBs are allocated and the second level of mapping is done by SDAP protocol from gNB to UEs. Transmission capabilities between gNB and UEs depend on how effectively radio resources are allocated to UEs [37]. Scheduling algorithms play a very vital role in allocating radio resources among UEs. Various resource allocation algorithms are proposed by researchers to compensate for channel loss and effectively utilize radio resources to satisfy QoS requirements [38]. Some of the classical resource algorithms are RR, PF, and Best CQI. RR algorithm is one of the simplest resource scheduling algorithms that works on a first come first serve basis and it allocates resources on equal probability to all the users without considering channel state information [39]. On the other hand, the PF algorithm [40] brings fairness to the users by assigning a priority preference. PF is the most widely used scheduling algorithm in the industry [41]. The best CQI algorithm selects the UE for scheduling the resources that have the highest CQI value; whereas, the UEs with a lowest CQI value will not get a chance to schedule the resources [42]. Based on QoS Guaranteed Resource Block Allocation (QGBRA) [43], (Best CQI higher deviation and the best CQI lowest algorithm were proposed which aim to achieve a good trade-off between throughput and fairness among users [44]. Other than the conventional scheduling algorithm, in the review of the literature section, we discuss various scheduling algorithms proposed by the researchers. Scheduling decision depends on various parameters, according to Ref. [45] there are various parameters are responsible for affecting the scheduling decision i.e., payload size, CQI, traffic types, HARQ, control channel overhead, and latency constraint [46].

2.6. 3GPP release for coexistence between eMBB and URLLC

The third-generation partnership project (3GPP) is a global collaboration of telecommunication standard organization which develop and standardize specifications for the mobile communication technologies. The main objective of 3GPP is to obtain interoperability and compatibility to diverse devices and communication components. For 5G radio access networks, the notable features and specifications are available in 3GGP releases viz. release 15 to 20. Releases 15–17 deals with standard 5G architecture, whereas advance 5G releases are discussed in Rel. 18–20 as depicted in Fig. 5. The subsequent sections will brief about the standard 5G specifications with reference to Rel. 15–17.

2.6.1. 3GPP release 15

3GPP release 15 is mainly focused on the eMBB services. It was the initial release for 5G NR which provides detail specifications about scalable numerology, flexible slot-based frame structure, MIMO technology, millimeter wave communication, etc. [26]. Instead of these services, release 15 focused on the coexistence mechanism of eMBB and URLLC. Here two main techniques were discussed on the coexistence mechanism viz. preemption (puncturing) and superposition [47]. In preemption, eMBB performance loss is higher because eMBB is unaware

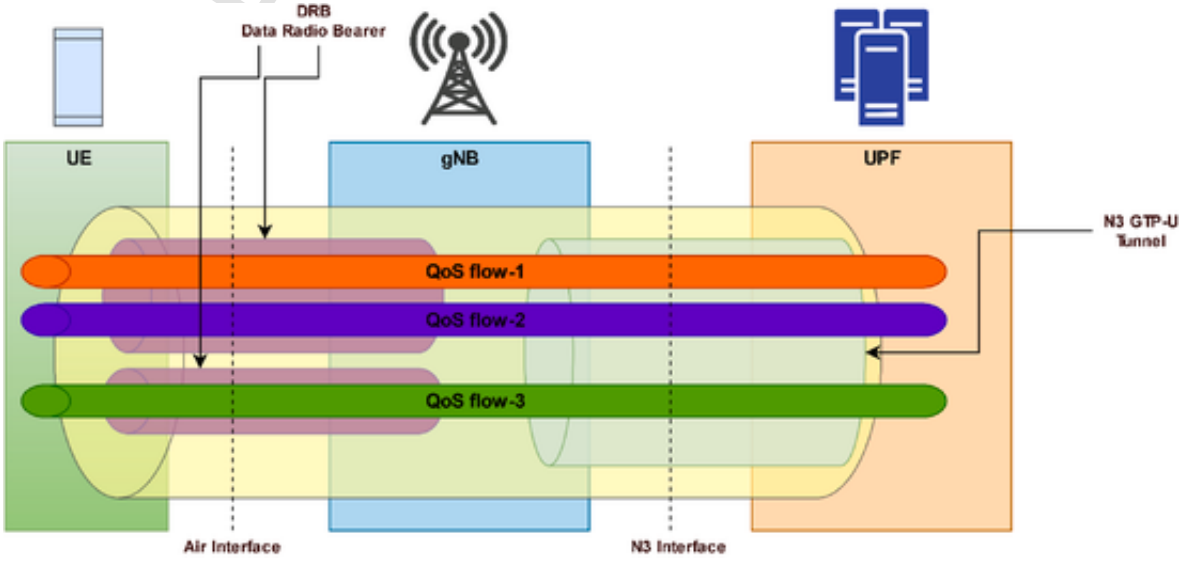

Figure: 4. QoS architecture for 5G NR.

Table 2
5G NR QoS parameters and 4G parameters.

Parameter	5G	4G
QoS Identifier	5G QI (QoS Identifier)	QCI (QoS class Identifier)
IP data Flow	QoS flow	EPC barrier
Flow/bearer Identifier	QFI (QoS identifier)	EBI (EPS bearer Id)
Reflective QoS	RQI (Reflective QoS id)	Not applicable
Data Session	PDU Session	PDN connection

of the punctured slots whenever a high amount of URLLC data arrived. Preemption technique is also discussed in section 2.3 of this paper. Superposition on the other hand combines and transmit both eMBB and URLLC signals concurrently. It utilizes two approaches viz. power domain superposition and time domain superposition. In power domain superposition technique, for transmission, the gNB combines both eMBB and URLLC signals at different power levels according to QoS requirements of the users. At receiving end, same power level differentiation is utilized by implementing advance signal processing technique to separate both type of signals. In time domain superposition, different time slots or subframes are allocated for eMMB and URLLC to ensure that both traffics can coexist without interference. Superposition technique significantly increases the eMBB performance as the URLLC load increases [48]. In addition to that, orthogonal scheduling technique adopted by 3GPP reserves certain number of resources in advance for URLLC services. Here basically two reservation techniques are utilized that are known as semi static reservation and dynamic resource reservation. In case of semi static reservation, gNB broadcasts both the frame structure configuration and frequency numerology. In opposite of that, in dynamic reservation technique, the frame structure information is updated by the control channel of the scheduled users. Dynamic reservation technique invents an additional overhead as compared to the semi-static technique. The major disadvantage of resource reservation technique is that the resources are wasted in case of non URLLC users [49].

2.6.2. 3GPP release 16

3GPP release 16 is mainly focused on the URLLC services and existing coexistence mechanism improvements that were targeted to reduce the latency and reliability requirements. It is also focused on the unlicensed spectrum(nr-U) utilization for the coexistence of "eMBB and URLLC services" with non 3GPP systems [50]. For end applications, Rel. 16 aimed on the industry automation especially transport industry including autonomous vehicles with stingiest latency of 0.5–1 ms [26]. It is also describing the coexistence of NB-IoT with NR which can utilize URLLC services and resource reservation technique to reserve certain amount of resource slots for URLLC services in case of puncturing. In addition to that, DL subcarrier puncturing approach was also introduced in Rel. 16. It also supports coexistence between LTE V2X and NR V2X that was deployed in Multi RAT environments. For flexible resource adaption, Rel. 16 introduced a cross link interference (CLI) handling technique to minimize the coexistence interference among eMBB and URLLC in LTE and NR environments [50]. Rel. 16 also worked on the network slicing in which eMMB and URLLC traffic can establish a coexistence in the form of virtual slice [51].

2.6.3. 3GPP release 17

3GPP Release 17 is focused towards communication on above 52.6 GHz licensed frequency band, and 60 GHz unlicensed frequency band. It also covers side-link and non-terrestrial access, enhancement of the drone communication technologies, and MTC for industrial sensors to reduce complexity and power consumption [52]. It also defines new and optimized QoS parameters for cloud gaming that require low latency and higher data rate services. In the case of multi access edge computing, an enhancement in optimization features was adopted for efficient mobility, discovery and positioning [51]. Rel. 17 also introduced network slicing phase-2 enhancement in which a standardized Slice/Service type (SST) value was proposed to establish a global interoperability for the slice, so that a roaming use cases can be improved [53]. In addition, release 17 employ an enhancement to semi-persistent scheduling algorithm for dynamic allocation of the resources in the coexistence scenarios. Here, gNB is responsible to preempt the transmission of PDSCH of one UE, if another UE is running latency critical applications [54]. No other coexistence approach was found in the release 17.

3. Classification of eMBB and URLLC coexistence approaches

We classified the coexistence of eMBB and URLLC is classified into five main categories as shown in Fig. 6. The subsequent sections will brief the work done by several researchers for coexistence of eMBB and URLLC.

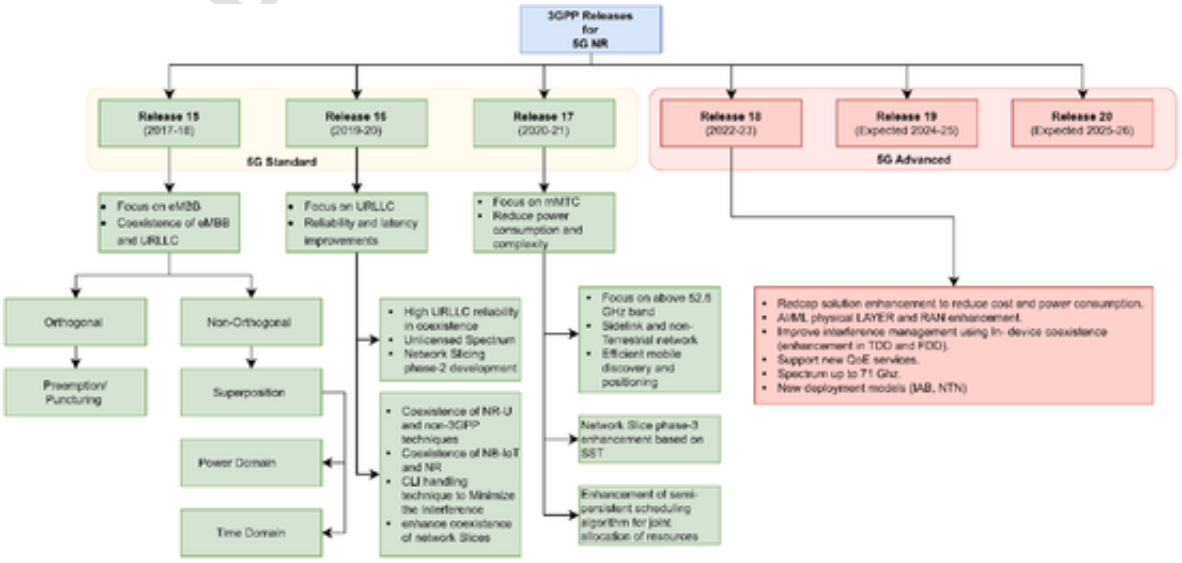

Fig. 5. 3GPP releases for 5G standard and 5G advanced.

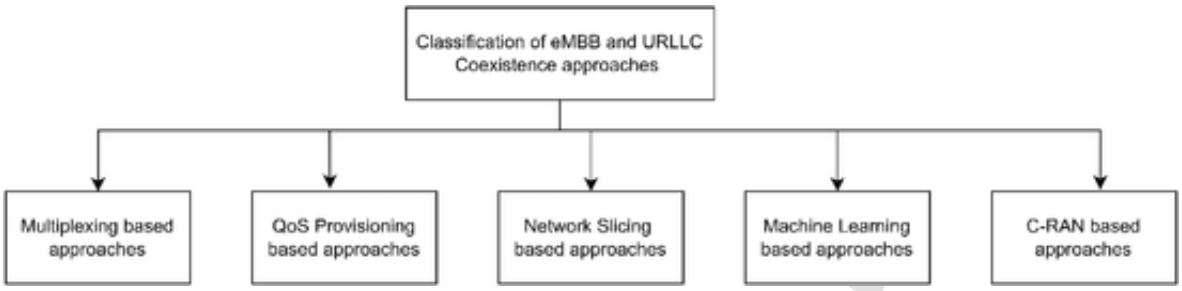

Fig. 6. Classification of eMBB and URLLC coexistence approaches.

3.1. Multiplexing-based approaches

This section provides a comprehensive literature based on the physical layer of 5G NR multiplexing schemes between eMBB and URLLC. In 5G NR multiplexing, the resource grids are combined in the time and frequency domain in the channel. For this, the gNB divides the channel into two types of slots i.e., slots and mini slots. eMBB traffic is being scheduled in time slots of 1 ms (15 kHz SCS) whereas URLLC traffic is scheduled in mini slots of the coexist region. Scheduling algorithms play a vital role in joint resource allocation in coexistence mechanism. In this regard, Anand et al. [55] presented a joint scheduling mechanism between eMBB and URLLC traffic, where URLLC traffic is placed in the mini slot and eMBB traffic is served in the given time slot. Authors exhibit and compare the rate loss function in linear, convex, and threshold models. The static scheduling supports non-opportunistic scheduling, in which the resource are equally shared among UEs. It was observed that in case of random puncturing each user faced an average of 50 % loss. The authors compare the opportunistic online gradient scheduler with the offline algorithms and found that the online gradient scheduler with minor modification puncturing and iterative allocation is optimal. The approach has benefits of efficient resource utilization which enhance the network capacity, also it has flexible adaption on the network. However, despite benefits, this approach has several limitations such as complexity of the algorithm, tradeoff between the prioritization of the traffic quite challenging and add additional overhead.

Efficient resource allocation depends on various parameters such as bandwidth, scheduling algorithm, signals quality and beamforming techniques. Bandwidth of the channel is one of important constraint for 5G deployment scenarios. Therefore, Han et al. [56] proposed a dynamic bandwidth part allocation scheme for 5G URLLC UAV with higher traffic rate. Here the authors discussed a dynamic multiplexing (DM) scheme that is a part of preemptive scheduling scheme. When URLLC traffic occupies a mini slot while puncturing the eMBB user traffic, some of the bits are erased in the process of puncturing, because DM has a limited error capability. But, on the other hand, Orthogonal Slicing (OS) provides better error handlining capability when eMBB bits are erased. The performance of the scheme also depends on the choice of modulation and coding scheme (MCS) used in both DM and OS. DM rely on the error correction capability of the Code Block Group (CBG)-based Max distance separable (MDS) code to compensate for the performance loss. If the URLLC load increases, OS outperform DM in terms of throughput. On the other hand, if the URLLC load is small, punctured bits can be corrected using MDS code. Therefore, a URLLC threshold parameter is required for deciding the bandwidth allocation for both the traffic. When the traffic reaches beyond the URLLC threshold, the system will select the OS scheme, otherwise it selects DM based scheme. The approach supports flexible utilization of the resources due to the bandwidth part. In addition, it also provisions low latency and high reliability requirements. However, if the traffic load increase instantly, the approach may face congestion problem.

On the other hand, NOMA mechanism is more effective as compared to OMA. The authors in Ref. [57] proposed an efficient coexistence approach between two traffics under Multiple Input Multiple Output (MIMO) - Non-Orthogonal multiple access (NOMA) system. NOMA is more friendly in resource allocation due to its successive interference cancellation algorithm at the receiver and it also increases spectral efficiency as compared to the Orthogonal Multiple Access (OMA) scheme. The authors formulate the objective function to maximize the eMBB system throughput in constraint of URLLC latency and proposed a joint user selection and power allocation approach. They adopted eMBB user clustering mechanism for not only stabling the structure of the eMBB traffic, but also to form a balance between the system performance and the computational complexity. They utilized the Gale-Shapley (GS) approach to formulate the user selection problem and “Successive Convex approximation with difference convex programming” (SCA-DC) algorithm for the power allocation problem. Based on SCA-DC, an iterative power allocation algorithm is proposed by authors and notified a significant improvement in the performance. The approach increases spectrum utilization due to NOMA but compatibility and interoperability on the existing infrastructure are quite challenging.

The complexity of the resource allocation algorithm is the main concerned in case of joint resource allocation in coexistence mechanism. Bairagi et al. [58] formulated the goal of maximizing the achievable rate of eMBB users concerning serving every URLLC request over a specific time. An optimization problem was formed based on these phenomena with the two constraints viz. one is related to serving every URLLC request on the given mini slot and other is related to every URLLC request that is served in the stipulated period. The problem was solved by preparing a preference list for the time slot t of eMBB users by the serving base station (SBS). Whenever a URLLC request is arrived, the SBS allocate resources based on the preference list that will satisfy the request of every URLLC user traffic. The proposed approach in this paper shows subsequent improvement in the achievable throughput and serving rate of URLLC user traffic as compared to the random approach. Bairagi et al. [47] extend their work as given in Ref. [58] with the same predefined objective to maximize the achievable bit rate of the eMBB user with the constraint of serving every URLLC traffic. The problem was decomposed into two parts, the scheduling problem of URLLC and eMBB users. A Penalty Successive Upper bound Minimization (PSUM) algorithm was used to solve the eMBB scheduling problem and an optimal transportation model was adopted to solve the URLLC scheduling problem. Furthermore, the first problem was redefined again and provided an algorithm based on PSUM and obtained a near-optimal solution. On the other hand, the second subproblem was redefined again and solved using a minimum cell cost algorithm and modified distribution. The authors also proposed a cost-effective heuristic algorithm to solve the first subproblem. The algorithm allocates resources to eMBB UEs in the rest of the time slot depending on how much proportional loss was accommodated during URLLC puncturing. The proposed approaches show an increase in the performance parameters as compared to mentioned benchmark algorithm. Due to this approach, subsequent improvements in eMBB loss and effectively schedule URLLC users were observed. On the other hand, due to the increased URLLC

load, the network performance reduces thereby not accommodating the requirements of all the users.

URLLC load increment in presence of eMBB users create a resource adjustment problem in the resource grid. Gerasin et al. [59] proposed a scheme for the division of the channel using the NOMA method for uplink transmission. They used a separate sub-band for eMBB traffic, and a common sub-band for both the eMBB and URLLC traffic that dynamically changes the width of the common sub-band. They identified the optimal parameters in the analytical modeling so that the URLLC traffic latency does not affect and increase the serving capacity of the eMBB traffic. The proposed scheme was compared with the OMA scheme and it significant increment in the parameter value was observed. The proposed scheme has added advantage of flexibility of choosing the resource grid to accommodate unpredicted URLLC traffic, but NOMA scheme itself invent additional overhead in the system. Additionally, if the URLLC load increases, the OMA scheme is preferred over NOMA. Tominaga et al. [60] discussed the scalable orthogonal and nonorthogonal slicing for the coexistence of the eMBB and URLLC traffic. With the use of NOMA technology, two traffics can simultaneously impose one another for sharing the bandwidth of the channel. In their scenario, they used NOMA technology with multiple URLLC users that supports Successive Interference Cancellation (SIC) and a frequency diversity mechanism as a solution for improving URLLC traffic. Initially, the URLLC signal is decoded first due to the strict requirement of latency, after that the successive interference cancellation mechanism will be achieved before decoding the eMBB signal so that the interference is not invented. Authors demonstrated following key aspects through Monte Carlo simulation: a) If the URLLC user has better channel conditions than eMBB users, the non-orthogonal slicing is an advantage over the orthogonal slicing for the sum rate of eMBB users; b) if the eMBB users have better channel condition than URLLC, the orthogonal slicing perform better than the non-orthogonal slicing for eMBB sum rate. c), in case of very high value of eMBB sum-rate as compared to URLLC, the non-orthogonal slicing is recommended. Here, frequency diversity is found to be the effective mechanism for accommodating both traffics. In case of equal load distribution to both type of traffics, the proposed approach does not consider the balanced parameters. On the other hand, it suffers from higher complexity and overhead issues due to utilization of both OMA and NOMA techniques.

The use of OMA and NOMA technique require extensive investigation in the field of energy efficiency. Sui et al. [61] discussed the resource allocation strategy in an energy-efficient environment for both eMBB and URLLC services for QoS satisfaction. Directly optimizing the energy efficiency in QoS requirements is a complicated task, hence they propose a sliding window-based resource allocation algorithm for optimizing the energy efficiency. The proposed algorithm reduces 16.7 % of power consumption and increases the energy efficiency by 23.9 % as compared to random resource allocation. The authors focused on energy efficiency requirements because these new defined systems consume a lot of energy. As far as the interoperability with the existing infrastructure is concerned, the proposed strategy may face compatibility issues.

Fairness is an important factor in the resource allocation strategy, thus authors in Ref. [62] proposed a fairness based distributed resource allocation (FDRA) algorithm to maximize the system throughput in terms of fairness, outage probability and channel reuse radius. The proposed algorithm performs better than the conventional algorithm in terms of system throughput with average outage probability constraints, and better fairness among small cell base stations. The approach provides a good fairness in the resource allocation, but it includes an additional overhead due to the additional signaling and communication between network entities.

A Heterogeneous network environment is a very challenging task for resource allocation due to different structures of various cells (femtocell, picocell, macrocell, microcell) and their different QoS requirements for the end-users [63]. It is also increasing the issue regarding fairness, interference management among various users. Here authors composed a NP-hard problem for joint resource allocation, inference minimization, and spectrum reuse maximization. They proposed a distributed randomized algorithm based on a probability-based heuristic that assigns PRBs to the users for resource allocation spectrum reuse maximization and interference minimization. Simulation demonstrates the validity of the algorithm in the given scenario. This approach provides equal resource allocation and the user satisfaction but due to the increase fairness, the throughput of the system decline. It is evident that the throughput and fairness are reciprocal to each other, the approach should take into account the threshold constraints between throughput and fairness.

Heterogeneous network environments efficiency could decrease due to lack of RAT selection technique. In Ref. [64], the radio resource management for multiple RAT selection, resource and power allocation, and traffic congestion control based stochastic optimization problems is formulated to maximize the network utility. To solve the optimization problem, the authors developed an online solution based on Lyapunov optimization as a joint problem where transmission rate is being controlled by the central controller and congestion policy is controlled by the end-user. They proposed a resource allocation algorithm for allocation of the resources and power allocation approach based on Lagrange duality approach and multiplier update technique. The proposed hybrid approach outperforms in terms of user throughput and RAT load. The approach has considered congestion, power allocation and centrally resource allocation solutions. The approach is based on Lyapunov optimization which is highly depends on both the accuracy of the system model and the parameters used in optimization.

The optimization algorithm in resource scheduling is require carefully planning in coexistence mechanism. Li et al. [65] proposed a heuristic SCA-RA (side-channel attack resource allocation) algorithm for URLLC and eMBB slicing in the 5G scenario. The main objective of the algorithm is to accommodate the maximum number of slices. Simulation results demonstrated that the proposed algorithm reduced the blocking probability of slice requests and the usage of the transponder and server. The approach was focused on resource allocation while considering the side channel attacks. However, there are other metrics such as fairness that needs to be considered during real-time deployments.

The real time deployment also needs to consider separate planning for resources scheduling in downlink and uplink. Therefore, Miuccio et al. [66] addressed the physical resource allocation problem in the 5G uplink massive machine type of communication, where massive devices generates small size of packets. The authors focused on PUSCH unused resources to mMTC devices that were failed their access attempts. The proposed PUSCH-based algorithm outperforms static and dynamic multiplexing algorithms in terms of successful communication and reducing cost with energy consumption. The proposed approach was focused on the uplink transmission and small size packets which are being generated by mMTC devices.

On the other hand, Wu et al. [67] presented a dynamic multiplex problem for URLLC and eMBB users in the downlink scenario where an indicator-free approach is utilized for resource scheduling. The authors proposed a correlation-based scheme that adds a correlation in the URLLC traffic before the transmission so that it can be easily identify whether the traffic is URLLC or eMBB. A precoding correlation scheme differentiates URLLC and eMBB traffic that can easily identify the punctured URLLC traffic and preempted slot for eMBB users. The proposed scheme confirms a better result and reduces computational complexity. In reality, it is quite challenging to predict whether the traffic is eMBB or URLLC. The approach includes an additional overhead in latency to determine the correlation between both the traffics.

To understand the correlation between the traffics require bit level mapping. A Trellis-Coded Modulation (TCM) using LDPC code-based technique was proposed in Ref. [68] for dynamic multiplexing of

URLLC and eMBB traffic. They considered two scenarios for modeling the code. In the first scenario both eMBB stream and URLLC stream considered one bit, whereas in the second scenario, two bits of eMBB, and one bit of URLLC is considered. It is observed that one-bit URLLC per symbol approach provides a guaranteed service delivery requirement for the URLLC stream. The proposed approach has better advantage in terms of URLLC requirements, but LDPC code require balance latency-reliability constraints. On the other hand, the design and implementation are quite challenging for real-time deployment.

The signaling optimization can reduce the overhead in the system, therefore, Chen et al. [69] discussed a new reference signal design for URLLC and eMBB multiplexing where flag signal was used a reference signal for the sharing of the resources. The amount of PDCCH monitoring was reduced by inventing a new reference signal and it was able to overcome the overhead by utilizing the spectrum efficiency. The proposed method in terms of BLER achieves an SNR of 3 dB as compared to the baseline method. The proposed approach reduces additional signaling to the system that help reducing overhead.

Signaling can be differentiate based on public and private network in 5G NR. Therefore, Yang et al. [70] discussed the multiplexing mechanism between URLLC and eMBB, where URLLC implementation was there in a non-public environment (local factory). They utilized the TDD approach for uplink and downlink channel resource scheduling. The major challenges faced by them were a high level of cross-link and a cross-interference. The problem can be solved by synchronizing the network for both the public and private network scenarios for eMBB and URLLC traffic environments. Factory environments utilizing both public and private network scenarios, in which interference between multiple radio node is a big problem, however effective beamforming and interference monitoring and management can improve the system capacity.

The joint implementation of use case scenarios of 5G NR is quite challenging, Chukwu et al. [71] discussed the 5G NR in different multiplexing usage scenarios such as mMTC and eMBB, and eMBB and URLLC. Simulation results demonstrated that the URLLC and eMBB have an average achievable channel capacity, whereas mMTC and eMBB have a greater capacity. It was observed that the data rate was higher and error rate was lower in mMTC and eMBB multiplexing due to larger bandwidth and millimeter-wave carrier application. The real-life implementation consisting of all three-use case scenario is challenging task, that requires proper radio frequencies and resource allocation planning. The resource grid also needs to be planned accordingly because of unpredictable mixed traffic of URLLC and eMBB.

eMBB service require higher data rate whereas URLLC requirements is low latency and higher reliability. Therefore [72], a time-division multiplexing approach was proposed for transmitting ultra-reliable low latency and higher priority data over the services. The proposed approach reduced the hybrid ARQ (HARQ) timing to half and maintained a vehicle to vehicle (V2V) latency goal of 99.99 %. The proposed approach provides guaranteed service level agreements and achieve desire latency requirement, but interference with the neighboring cell while transmitting signals in the adjacent time slot, thereby offering limited flexibility in the changing traffic conditions. Table 3 shows the findings of a few multiplexing-based approaches.

Table 3
Multiplexing-based approaches.

Ref	Problem Formulation	Techniques	Findings	Remarks
[55]	eMBB maximization with a priority of URLLC with linear, convex, and threshold models	Joint Scheduling based on an online gradient scheduler	The larger value of δ (time slot) limits URLLC traffic and maximizes eMBB throughput	Fairness policy and QoS enable policy
[47]	Maximization of minimum expected achieved rate (MEAR) of eMBB UEs along with URLLC traffic	Penalty successive upper bound minimization (PSUM), transportation model (TM), Heuristic algorithm, minimum cell cost (MCC) and modified distribution (MODI)	Average MEAR increased as compared to RS, EDS, MBS, PS, and MUPS	Power level put to zero to avoid interference during puncturing of slots but it can produce an extra delay
[73]	Optimization of Spectral efficiency and URLLC latency	Multi-user preemptive scheduling (MUPS)	Gain in average cell throughput	Performance of eMBB users is good until the interference reference is dominant
[74]	Maximization of eMBB rate and minimization of URLLC rate	Null space-based preemptive scheduler (NSBPS)	Instant scheduling for sporadic URLLC traffic with minimum impact on overall ergodic capacity; URLLC traffic does not not experience queuing delay; safeguard for URLLC traffic for possible interference	Bigger antenna grid is required for higher performance
[57]	Maximize the data rate of eMBB with constraints to URLLC latency; Joint user selection and power allocation.	eMBB user clustering method, Gale Shapley (GS) matching process, and SCA-DC	MIMO-OMA provides higher throughput per eMBB user but lowers in entire throughput; eMBB performance decreased with the increased number of URLLC users; fairness increase	Small clusters of eMBB users were used for demonstrating the problem
[75]	Problem is redefined based on the effective capacity (EC) model; frequency power allocation problem is defined as a mixed-integer nonlinear programming problem	Two-step optimization problem: slack the frequency variable to optimize the frequency allocation and then decomposing the function using difference of convex (DC) algorithm	The proposed scheme offers 12 % higher throughput and 49 % higher gain.	The delay of URLLC traffic is higher

(*continued on next page*)

3.1.1. Discussion

As mentioned earlier, the multiplexing approach combines both frequency domain and time domain at the physical layer of 5G NR. Here, the major challenge resides in managing the interference among coexistence of eMBB and URLLC. To decrease the mutual interference between these services and improve overall system performance, techniques such as interference coordination, interference avoidance, and interference cancellation should be considered. As per 3GPP specifications, the coexistence of eMBB and URLLC is supported by puncturing and super-positioning techniques. The performance of puncturing

Table 3 (*continued*)

Ref	Problem Formulation	Techniques	Findings	Remarks
[76]	Joint multiplexing technique used in a vehicular network environment where vehicle node and roadside unit (RSU) node modeled in 1D position	To guarantee the reliability of the URLLC users, guard zones are deployed around the vehicle receiver; eMBB transmission inside the guard zone was prohibited	URLLC users' traffic is well protected due to the proposed guard zone technique that leads to an increase in performance as compared to without guard zone scheduling; better rate coverage	eMBB users' performance parameters need to be considered as the guard zone has restricted permission
[77]	Resource allocation problem formulated as a sum-rate maximization problem	Proposed an optimization-based scheduling algorithm and a heuristic algorithm	The performance of optimization-based scheduling algorithm is found better than heuristic algorithm	If the no of URLLC users increases it is difficult to get the minimum sum rate for the eMBB users
[78]	Cellular Device association and power control formulated as Non-Convex Optimization	Framework for maximizing the energy efficiency of NOMA; User association problem solved using dual theory approach and optimal power control problem resolved using new sequential quadratic programming (SQP) Technique	The proposed algorithm outperforms in terms of energy efficacy as compared to the benchmarked algorithm; it provides min rate for each user, hence satisfying QoS requirements and successfully decoding SIC	Framework needs to consider imperfect channel conditions
[79]	Cell admission control (CAC) problem formulated as a minimization problem for eMBB and URLLC users	Sequential convex programming (SCP) to find the suboptimal solution of the CAC problem	The CAC algorithm's main goal was to admit the largest number of eMBB users to satisfy the URLLC users' constraints	Coexistence of eMBB and URLLC must be done without reducing the QoS
[65]	Side Channel Attack (SCA) affects the resource allocation strategy	SCA-RA heuristic algorithm	The blocking ratio was reduced	AI techniques can improve security
[80]	Resource allocation problem is formulated as non-convex, mixed-integer, multi-objective optimization	A novel hybrid approach based on puncturing and superposition	The proposed approach is beneficial in URLLC admission control and scheduling	The better trade-off between eMBB throughput and admission of URLLC
[81]	Joint Resource allocation Problem	Noma super positioning and puncturing technique	The proposed approach shown a significant performance improvement	Spectral efficiency and fairness depend on tuning of parameters
[82]	Jointly optimizing the reflection coefficient of the reconfigurable intelligent surface (RIS) elements, power allocation, and RB allocation	RIS-aided THz communication system utilized using deep and ensemble learning method	Improved spectrum efficiency and satisfied eMBB and URLLC service requirements	Ensemble learning performs real-time resource management
[83]	Pre-emptive priority service to the base station	Priority-based implementation of eMBB and URLLC coexistence	Preemptive priority based service isolates eMBB and URLLC traffic	The number of antennas impacts the performance

scheme is found to be better than the superposition technique due to better handling of interference and improving resource scheduling in terms of throughput and latency. eMBB throughput depends on the number of slots punctured and the amount of URLLC load, whereas URLLC latency depends on the URLLC payload size and packet arrival rate. While puncturing, the power level of eMBB users need to be set to zero to reduce the interference, which results in reduced performance of eMBB throughput. In that case, the NOMA technique (superposition) is found to be more effective as compared to the OMA technique, because, it places resources in a non-orthogonal fashion. In addition, due to the SIC algorithm in NOMA, the interference is cancelled at the receiver end.

Based on literature, it can be stated that the sporadic nature of URLLC is another major gap in the coexistence. Many approaches were adopted by the researchers (as discuss in this section) to enhance the eMBB throughput while maintaining URLLC reliability. Some of them were based on the admission control strategy in case of higher traffic load, whereas some of them were based on effective interference cancelation, and joint resource scheduling. In addition, the prediction of URLLC traffic can help supporting the eMBB and URLLC coexistence to a great extent. If the gNB already prepared a cluster for the eMBB and URLLC users, radio resources will be optimized. According to the preference list in the given cluster, users will get the resources as per the requirements. This approach is found to be an effective approach if both the channel allocation and resource allocation rules are predefined. Some of the realistic scenario which require multiplexing of eMBB and URLLC services are industry automation in which eMBB provide video surveillance and remote monitoring. URLLC services on the other hand utilizes critical industry automation applications such as control of robotic arm, emergency shut down, etc.

3.2. QoS provisioning-based approaches

QoS provisioning-based approach deals with provisioning of QoS parameters among eMBB and URLLC co-existence mechanism. Most of the work focused on the QoS parameters improvement using eMBB and URLLC coexistence mechanisms. The performance of the well-known Round Robin (RR) resource scheduling algorithm is found to be satisfactory because it does not takes into account the channel quality indicators [84]. To obtain higher performance, a scheduling strategy should consider the channel quality indicators especially for GBR type of services. Therefore, the authors purposed an enhanced joint scheduling (eJS) algorithm that utilize the features of best channel quality indicators (CQI) higher deviation and Best CQI lower deviation. The main aim was to improve the fair allocation of the resources among UEs so that another UE is not penalized due to that resource allocation strategy. The fairness of the resource allocation is calculated based on the Jain fairness index. The authors formulated an optimal resource allocation problem between the GBR and non-GBR Data Radio Barriers (DRBs). The problem was solved using the eJS algorithm utilizes a pseudo heuristic approach to maximize the system throughput in terms of reaching the fairness level among DRBs. The proposed algorithm is based on FDD communication that effectively utilizes QoS parameters for optimal resource allocation.

QoS is the important parameter to meet the user demand and provide services as per user experience services [85]. Due to the complex designing structure of the millimeter (MM) wave technology, the han-

dling of non-asymptotic error structure regime is very challenging for providing end-user latency. To meet these challenges authors proposed a statistical QoS-driven resource allocation policy using millimeter wave technology for both asymptotic and non-asymptotic regime. The Shannon capacity formula is inefficient for achieving maximum data rate due to the block error rate probability introduced by blocklength coding in the asymptotic scheme. The proposed technique outperforms the baseline approach in terms of maximum achievable rate of the end users in satisfying QoS requirements. The approach has an additional advantage of QoS satisfaction of the users, but statistical resource allocation approach only focused on the statistical characteristics of the traffic rather than deterministic QoS guarantee. Despite these, millimeter waves have sever problems such as pathloss that needs to carefully plan. Korrai et al. [86] firstly formulated an optimization problem with joint resource allocation for RBs followed by formulation of power allocation that satisfies SNR, latency, and isolation constraints. A successive convex approximation algorithm is proposed for joint resource allocation for RBs and power allocation. The proposed technique achieves higher performance in terms of eMBB data rate, URLLC latency, mMTC queue with low power consumption. The proposed approach supports adapted numerology and analyzes the impact of imperfect channel condition. The approach is sensitive to channel variation, that needs to focus the robustness of the algorithm in channel variation conditions including fading and mobility.

On the other hand, Zarin and Agarwal [87] focused on joint radio resource allocation for multihoming calls in heterogeneous environments. The main objective of their optimization problem was to maximize the system sum throughput. They formulated an optimal subcarrier and power allocation-based approach from the OFDM-based network and timeshare allocation from WLAN. Simulation demonstrated that the proposed algorithm provides a faster convergence rate and a higher total system sum-rate. The proposed approach provides effective resource utilization and enhances QoS efficiency, but in realism it suffers from interoperability as it operates on different networks, standards, and protocols. In addition, getting fairness among users in heterogeneous environments is also quite challenging.

Instead of throughput improvement, fairness among user is also essential in joint resource scheduling. In this regard, Panno and Riolo [44] proposed QoS service-aware resource allocation schemes that are used for a tradeoff between throughput and fairness among the user. The authors proposed two approaches to effective utilization of resources 1) Best CQI highest deviation 2) Best CQI lowest second low. The first approach was concerned about maximizing the throughput whereas the second approach was concerned about maximizing the fairness among users. The results of their proposed approach were compared with the benchmark approaches in different channel conditions and traffic capacity. The proposed scheme provides better performance as compared to other algorithms under consideration in terms of throughput and fairness among UEs. It also need to carefully analyze the tradeoff parameters, because both throughput and fairness are reciprocal to each other.

On the other hand, Gharam and Boudriga [88] proposed a game-theoretical model based on different types of network technology provided by different network operators. They proved the convergence of the game theory based on AP and UE with two different utility functions. Simulation shows the effectiveness of the algorithm in terms of system performance. Game theory approach provides a significant improvement in the resource allocation due to the self-adaptation and organization. While deploying, the proposed approach needs to consider several challenges such as complexity, information assumption and coordination among entities.

For joint resource allocation in support to coexistence, the QoS framework of 5G NR require subsequent improvements. In this regard, Miuccio and Panno [89] proposed a framework relate to the QoS aware resource allocation scheme for GBR and Non-GBR services. The algorithms work on two levels: the first level is about channel aware resource allocation for multiple numerologies (CARAM) and the second level defines the same priority with the multiple numerologies to satisfied the fairness among resources. The overall objective is to maximize both the number of services for GBR and the system throughput. The proposed framework provides a QoS satisfaction rate among satisfied GBR services that lead to improved system throughput and increased spectrum efficiency, with fairness constraints and priority levels. The proposed scheme provides QoS guarantee for both GBR and Non-GBR services. Multiple numerology also provides flexibility to choose the resource grid for scheduling of the resources.

In recent era, TDD environment is more popular for joint resource allocation and QoS satisfaction of the users. Therefore, Esswie and Pedersen [90] proposed a QoS-aware TDD system framework for industrial factory development in the 5G network. Initially they choose dynamic TDD mode with optimized uplink power control setting and QoS aware dynamic user scheduling for TDD link selection. Finally, a reinforcement learning-based approach for selection of base station specific TDD frame configuration for different load regions was introduced. The proposed scheme shows a 68 % URLLC outage latency reduction requirement as compared to a non-QoS-based scheme. The proposed scheme is complex in the real implementation scenario because of varying load of base station according to the usages in industrial environments.

Channel Quality Indicator and monitoring procedure is an important method to gain visibility of the channel quality among individual users. In this regard, an enhancement channel quality indication measurement and monitoring technique was proposed by the Poccovi G. et al. [91]. Firstly, to estimate the signal-to-interference-noise-ratio (SINR), authors collected the channel quality samples at the user equipment. Secondly a new CQI reporting format was proposed to better guide the downlink scheduling and link adaption decision for URLLC traffic. Simulation demonstrates that the proposed technique outperforms the traditional CQI measurement and reporting scheme in terms of latency. The proposed scheme provide lead to traditional CQI measurements in terms of improved link adaption and QoS guarantee for the URLLC users. However, by doing this, it may introduce feedback overhead in term of bandwidth and signaling.

Resources are multiplexed in the time and frequency domain in two different ways viz. orthogonal and non-orthogonal. In this regard, Akhtar and Arslan [92] modeled a packet scheduling and resource allocation problem in multi numerology system. They proposed an adaptive multi-numerology resources allocation algorithm having the capability of multiplexing different numerology and provide effective resource allocation. Simulation demonstrates that the proposed algorithm has higher throughput in different traffic with different QoS requirements. The proposed approach provides comprehensive coverage and QoS aware packet scheduling. However due to interoperability and compatibility issues, the approach may face difficulties during real world deployments.

On the other hand, Gerasin et al. [93], discussed the joint resource allocation problem for eMBB and URLLC multiplexing with flexible-NOMA in the uplink environment. They proposed a resource scheduling and power allocation algorithm for eMBB users with strict latency constraints for URLLC users that can adaptively tune MCS for URLLC transmission. The proposed algorithm shows significant improvement in the eMBB system throughput with the constraint imposed by grant-free URLLC transmission. Further, it also reduced protocol overhead that was invented during the reconfiguration of grant-free URLLC transmission. The grant free transmission mechanism reduces the transmission latency for the URLLC services and enhance the eMBB throughput. It also needs to consider and mitigate the interference from both traffics.

In a grant-free (GF) request, straightly data is transferred to particular UEs without queue, but if these resources are not utilized by URLLC traffic, gNB schedules these resources for eMBB traffic for better utilization of the spectrum efficiency [94]. For this situation, there is a possi-

bility of collision because some of the URLLC traffic is still active, and these GF resources overlap with eMBB and URLLC traffic. They proposed a two-step overlapping indication and hybrid automatic repeat request (HARQ) feedback mechanism to improve the URLLC and eMBB multiplexing. Simulation results indicated an improvement in the spectrum efficiency and reduced error probability due to the demodulation reference signal miss detection (DMRS). This approach carefully mitigates the interference from both type of traffics which improves the system performance in terms of error decoding capabilities and spectral efficiency.

Computational resources also equally important in case of joint resource allocation. Therefore Jankovic et al. [95] proposed a joint computational and radio resource allocation framework that works on analyzing the performance of individual services of the users based on QoS parameters. The authors proposed a computational load distribution algorithm that balanced the workload among users. The simulation results demonstrated that the proposed solution reduced packet drop ratio up to 15 % and user data rate up to 7 %. In addition, the resource granularity problem was resolved by adapting the allocation interval. The study was concerned to radio resource differentiation of various traffics as per QoS requirements. The approach needs to carefully analyze the practical deployments.

Integration of the simulation scenario in the 5G ecosystem to boost eMBB service in a highly dynamic scenario that is the terrestrial network infrastructure environment. The authors in Ref. [103] proposed an optimization framework that can exploit available resources in the given network slices to meet the QoS requirement for different users. They also suggest that the neural network-based solution provides effective optimization of the network resources. Table 4 shows the findings of a few QoS provisioning-based approaches. The framework provides an enhancement to eMBB capabilities in terms of throughput, latency, and reliability for satellite communication. The authors also considered cross layer optimization to enhance the QoS for eMBB users. However, one of the major challenges lies in integration of 5G and satellite communication due to coordination, compatibility, and standardization of the different protocols and network components.

3.2.1. Discussion

5G NR provides a separate framework for adoption of QoS requirements (i.e., latency, throughput, reliability, capacity, and mobility) to support a wide range of applications and services. However due to the adaptable environment, the mobile operator can generously set their own QoS requirements. As stated earlier, 5G NR supports three types of resources viz. GBR, Non-GBR and delay critical GBR. QoS parameters of GBR and delay critical GBR can be mapped according to the services of URLLC, whereas GBR and non-GBR resource can be characterize through eMBB services. In this regard, resource partitioning must be critical to allocate the resource according to the time and frequency domain in the resource grid. Here, the role of scheduling algorithm becomes crucial in scheduling the resources as per both QoS framework policy and user requirements. Several research were carried out towards designing of joint resource allocation algorithms which not only tackle the coexistence, but also map resources according to QoS requirements. Fairness on the other hand is one of the important metrics that needs to be considered while designing the QoS aware resource scheduling algorithms. Additionally, a proper admission control strategy also results in enhanced performance of eMBB and URLLC coexistence. It is observed that the joint resource and power allocation faces complexity issue while scheduling of the resources. Here, the inclusion of a heuristic based approach can significantly optimize the resources in both time and power domain. In addition, several other factors such as interference management, traffic shaping, flexibility, adaptability for different deployment scenario, and QoS prioritization should be carefully planned.

Table 4
QoS provisioning based approaches.

Ref	Problem Formulation	Technique	Key Finding	Remarks
[96]	Weight sum throughput of the URLLC users with weight coefficient of each user for QoS requirement	A Novel multi-objective resource allocation scheme for eMBB and URLLC users	QoS requirement satisfied with higher number of URLLC users; due to puncturing, average sum rate of URLLC users decreased	The average sum rate of eMBB and URLLC users depend on the value of parameter β
[97]	Effects of the joint allocation of the resource on the spectrum sharing approach, and operator's independence and fairness	Fully Hybrid spectrum sharing (FHSS) technique	Improvements in the outage probability; enhanced system performance in terms of QoE, load balancing, signal quality maximization, and a higher degree of fairness (96.08 %)	Efficiency of the proposed approach was found better in terms of both operators' independence and fairness
[32]	Mapping heterogeneous flow from multiple users and transport block	Configuration-based Assignment and packing algorithm (CBAP)	3 % overallocation occurs in transport block in more than a hundred flow entries	Dynamic allocation approach was followed instead of puncturing
[98]	Due to mixed traffic (eMBB, URLLC, mMTC) efficient scheduling technique is required for met predefined QoS for each slice	Priority-based polling with multiple scheduling schemes for different data traffic	Priority-based polling scheme showed a better utilization factor and efficiently allocate resources, especially for URLLC slices. It also satisfied the QoS for the particular user in a C-RAN architecture	Analysis of cyclic and random polling for multiple 5G use cases was achieved
[99]	Prioritize GBR user QoS admission to Non-GBR service	Admission Control L3 Resource allocation L2	model's appropriateness has been confirmed with low percentage errors (3 %)	Joint Optimization of the performance of the different tenants
[86]	Joint RB and Power allocation	A successive convex approximation algorithm	eMBB data rate increased; URLLC latency satisfied; Low power consumption	Power minimization by considering mixed numerology-based frame structures
[89]	Maximization of GBR users with Priority Constraints and maximizing system throughput with GBR services	Channel-Aware Resource Allocation for Multi-numerology (CARAM)	Increased System Throughput, number of satisfied GBR services on higher priority was found to be higher as compared to baseline algorithms	Large GBR services satisfied with high priority
[90]	TDD frame selection and resource scheduling	TDD frame selection framework based on reinforcement learning	Improvement in URLLC latency	Carefully modeling of learning objectives is required
[92]	Resource allocation problem for the non-orthogonal multi-numerology system	Adaptive Numerology Resource Allocation (ANRA) algorithm	Increased system throughput and fairness among resource allocation	Increase number of UEs can saturate the system throughput

(continued on next page)

Table 4 (*continued*)

Ref	Problem Formulation	Technique	Key Finding	Remarks
[93]	Multiplexed different QoS flow with requirements	Flexible NOMA	Reduced protocol overhead in grant-free URLLC transmission; increased throughput of eMBB	The scenario can be used for mobile URLLC users
[95]	Joint Computation and radio resource allocation	Computation and Backhaul radio resource allocation algorithm	15 % lower eMBB packet loss ratio, i.e., 7 % higher eMBB user data rate	Bursty eMBB traffic faces higher packet loss
[100]	New Resource allocation scheme model	Priority-based load-adaptive preamble separation (PLPS) scheme	Achieve different QoS requirements of eMBB, URLLC, mMTC	Random access channel throughput increased
[101]	Novel proportional fairness-based resource allocation problem for coexistence mechanism	QoS guaranteed resource allocation framework	Met latency and reliability of URLLC and maximize the eMBB performance with fairness.	The approach can be applied in TDD environment for better results
[102]	The joint optimization problem for bandwidth and power allocation	Dynamic resource allocation and puncturing strategy	Achieved low latency for URLLC service and minimize the eMBB service loss	Computation complexity of the algorithm is lower than the baseline heuristic algorithm

3.3. Network slicing-based approaches

Network slicing is an effective approach for network function virtualization (NFV). In slicing, eMBB and URLLC traffic frame into slices and services will be offered as per the requirement of UEs. Flexibility and scalability can be achieved using the services of virtualization in the coexistence mechanism between eMBB and URLLC. This section presents an extensive survey on the coexistence mechanism between eMBB and URLLC using network slicing.

Coexistence mechanism require an optimization in case of resource joint resource allocation. Therefore, Song et al. [104] proposed an optimization problem as subchannel allocation and power control as an infinite-horizon average-reward Constrained Markov decision process (CMDP) problem. With the use of this algorithm, an optimal policy is derived, but this optimal policy faced a serious dimensionality problem that can be solved by dynamic programming. In addition, an online stochastic learning algorithm was proposed to solve the subchannel allocation problem. The proposed algorithm outperforms as compared to the baseline algorithms in terms of converging rate and user performance. The proposed approach provides adaptive resource management due to virtualization and considers dynamic nature of network conditions and user demands. For deployments, the algorithm requires adequate processing power. Instead of processing power, service quality cost is equally important to achieve desire QoS satisfaction. The authors in Ref. [105] formulated an optimization problem which ensure latency and reliability requirements of the URLLC services while improving the QoS for eMBB services. The dynamic optimization model is utilized for power allocation and service quality as a cost optimization function with latency constraints. Lyapunov optimization is designed for long time scale bandwidth allocation and short time scale service control. The proposed algorithm outperforms three baseline algorithms in terms of hard latency, power consumption, and total cost. The proposed approach provides QoS service provisioning and efficient network slicing among users as per demands, however it requires additional signaling and controlling mechanisms to adjust the real time deployments.

On the other hand, Oladejo and Falowo [106] addressed the latency-aware dynamic resource allocation problem by formulating it as a maximum utility optimization problem in the 5G Heterogeneous environments. They proposed a Genetic Algorithm (GA) enabled intelligent latency aware based dynamic resource allocation scheme. Here different network slices i.e., eMBB, mMTC, URLLC were considered for the assignment of the resources. The proposed approach outperforms as compared with other baseline approaches such as static slicing resource allocation and optimal resource allocation approach. The proposed approach supports slicing, multitenancy, and effective scheduling of the resources. However, for real world implements, the proposed approach may suffer from higher computational complexities due to stochastic nature of GA.

To meet the need of every slice for resource allocation in QoS constraint is a challenging task. In this regard Fossati et al. [107] proposed a joint resource allocation problem by proposing a versatile framework called MURANES that was based on order weight average operator. The goal of the framework was to allocate resources to every slice in a balanced manner to satisfy the QoS requirement for each user. Framework analyzed many important findings such as it was not allocating unwanted and surplus resources, allowing ideal capacity to support traffic peak and not all resources are equally congested. The framework suggested the tradeoff between the resource utilization, slices separation and service performance. It provides an efficient resource allocation solution for complex resource constraints but limited to service level agreements constraint of non-linear resources.

Efficient resource allocation also requires channel capacity analysis and slice separation. Therefore, Santos et al. [108] proposed a Max-Matching Diversity (MMD) algorithm for effectively allocating the channel for eMBB with the consideration of H-OMA and H-NOMA slices. The proposed technique provides higher efficiency in terms of eMBB achievable rate and URLLC reliability. The proposed scheme increases frequency diversity which lead to proper accommodation of the user in channel, but whenever the SIC algorithm fails to decode the URLLC, a significant performance loss can be observed.

Hossain and Ansari [109] formulated a joint power and bandwidth allocation problem aimed to maximize the throughput by offering a scheduling preference to higher priority slices. The problem is defined as a mixed-integer non-linear programming (MINLP) problem considering the Channel State Information (CSI) of each user allocating the resource in frequency and time domain. The MINLP problem is relaxed and converted the same into a convex problem. The authors also find out that the convex approximation is very near optimal by solving the MINLP and convex problems. Simulation demonstrated that the QoS requirement of all the user slices was satisfied and there was a balance in throughput maximization with the tradeoff fairness. The proposed scheme with the extensive solution shows the real-life deployment scenarios. Whenever the network load increase beyond the limit, the scheduler diverts the resources toward best effort slice instead of GBR.

To effectively allocate dynamic resources, Chi et al. [110] proposed a new random compensation service scheme based on the statistical characteristics of arrival flow and fading channel slice adjustment factor. To achieve a guaranteed QoS requirement, they utilized the effective bandwidth/capacity theory by proposing a two-hole leaky bucket random service mechanism. The proposed scheme is suitable for communicating among users in a business scenario using video calls. The scheme has benefits of managing strong heterogeneous traffic flow such as voice and video traffics. The slice adjustment factor in this approach supports simplicity and can be considered for initial deployment. On the other hand, in case of both burst traffic and low latency constraints, the proposed scheme consumes more bandwidth that leads to significant filth in the overall performance.

Shi et al. [111] proposed a mechanism to determine the number of radio resources required for the eMBB service and analyzed the delay probability of the URLLC services. The optimization problem is formu-

lated based on a network resource pre-allocation scheme. To solve the optimization problem, a low complexity heuristic algorithm was proposed by authors. The simulation result demonstrated that the proposed soft slicing scheme achieved higher resource utilization efficiency and strict QoS requirement for eMBB and URLLC users. The proposed scheme provides added benefits of collision free inter-gNB resource sharing and network level RB allocation, which utilize service isolation and resource multiplexing. However, in this approach prediction of the URLLC traffic is not properly characterized.

Spectrum efficiency and URLLC reliability achievement is the main key problem in the 5G network. Therefore Ma et al. [112] adjusted the objective function of the optimization problem and formulated the problem as a convex optimization problem. To find out the optimal solution of resource allocation problem, they proposed a slice resource allocation algorithm based on the Powell Hestenes Rockfeller (PHR) method, and branch and bound technique. The proposed algorithm provides higher spectral efficiency and URLLC reliability as compared to baseline algorithms. The proposed algorithm was tested in the low bandwidth and low load conditions, which may be inefficient for higher number of URLLC users. In addition to that, the approach was utilized in a single RAN environment which is another barrier for real life deployments.

To meet the QoS requirements for the flexible traffic, 5G utilizes MEC technology in the NFV environment with the help of network slices [118]. Dynamic characteristics of the resource allocation can provide an uncertain real-time resource requirement. Here authors formulated an optimization problem of dynamic end-to-end resource allocation in MEC environment using the Markov Decision Process (MDP). Their goal was not only to maximize the resource efficiency but also to satisfy the QoS of each slice. Based on a policy gradient-based "Proximal Policy Optimization" (PPO) algorithm, they proposed a solution to this problem by developing joint learning algorithm known as "Independent Cooperative PPO-based Resource allocation" (ICPRA) and "Jointly Cooperative PPO-based resource allocation" (JCPRA). The algorithm performed better than other baseline algorithms and provided a balanced resource allocation to the users. It is evident that the complexity of the network slice is one of the main challenges that needs to be consider while working on the resource allocation using slicing. The authors provided an effective solution for the resource allocation in network slicing environment that help setting important baseline for practical development especially slicing.

On the other hand, Gonçalves et al. [119] presented a blockchain based approach for distributed network slicing for e-health environments. The experimental work was performed on the behavioral analysis of TCP/UDP protocols while doing the traffic prioritization of the slices. The authors gave important guideline in the form of architectural design of network slicing for hospital environments by utilizing blockchain technology to design a secure environment. However, inclusion of both public and private network scenarios can bring significant gain to the overall approach. Based on literature analysis, the key finding of network slicing based approaches are presented in Table 5.

3.3.1. Discussion

Network slicing in 5G NR is a powerful method to offer customized services, enabling vertical sector applications, and utilizing network resources efficiently. It enables operators to meet the varying needs of various industries, use cases, and applications while optimizing resource allocation and guaranteeing an excellent user experience. Network slices are helpful in framing different traffic prioritizations of eMBB and URLLC slices to form a virtual network resource environment. Effective network slicing improves user service satisfaction levels and resource allocation among users. There were various techniques discussed in the research literature where some of the major work was focused on online scholastic algorithm, Lyapunov optimization, genetic algorithm, MURANES based scheme for considering weight factor, max

Table 5
Network Slicing based approaches.

Ref	Problem Formulation	Techniques	Key Finding	Remarks
[56]	Bandwidth part allocation problem formulated using dynamic multiplexing and orthogonal slicing	Preemptive based scheduling	If URLLC load is high than OS outperforms DM; if URLLC load is small punctured, bits can be corrected using MDS code	DM relies on the error correction capability of the CBG based MDS code to compensate for the losses
[104]	Subchannel allocation and power allocation as a Constrained Markov decision process (CMDP) problem	Distributed Online learning algorithm.	Improved convergence rate and user performance	Correct state information is not available
[113]	Radio resource scheduling for mixed traffic and QoS requirements for each user	A heuristic algorithm (dynamic slice resource provisioning) for dynamic mixed slicing; an intra-slice shape-based algorithm to boost the QoS fulfillment to the user	The proposed scheme outperforms NVS and NetShare algorithms for resource utilization and average slice satisfaction ratio	Queue length is higher in the case of the URLLC slice
[114]	Dynamic resource sharing	Share Constrained Proportionally Fair (SCPF) algorithm	Higher performance gain during imbalanced load of a slice	Performance maximization via admission control
[60]	Jointly allocation of the resources in network slicing-based environment	NOMA is used to improve the number of URLLC users using network slicing mechanism in orthogonal and non-orthogonal with eMBB users	If URLLC users have better channel conditions than non-orthogonal slicing is an advantage over the orthogonal slicing for the sum rate of eMBB users and vice-versa	Complexity computations can also be considered for enhanced performance
[105]	URLLC power and Latency constraint improvement in QoS for eMBB users	Optimal Network Utility Algorithm based on Lyapunov optimization	The proposed algorithm outperforms in hard latency and power consumption as compared to the other benchmarked algorithms	Hard latency is inversely proportional to power consumption
[115]	Personalized service preferences and evolutionary interest relationships to model the complex and dynamic network environment	propose a bio-inspired virtual resource allocation scheme with slice characteristic perception (BVRA-SCP) for 5G-enabled IoT networks	The proposed algorithm provides flexibility in the dynamic resource allocation and optimization of network slices as compared to the traditional wireless network resource allocation scheme	The work can be improved to add more user's characteristics related to the social behaviors

(continued on next page)

Table 5 (*continued*)

Ref	Problem Formulation	Techniques	Key Finding	Remarks
[108]	Radio resource sharing between eMBB and URLLC	Max-Matching Diversity Algorithm with H-OMA and H-NOMA slices	NOMA-MMD outperforms during small values of URLLC rate, if the rate increases OMA-MMD performs better	Complexity increased to execute the channel allocation
[116]	Resource allocation optimization	Joint intelligent traffic prediction and radio resource management framework using LSTM	Improvement in resource utilization by adapting dynamic traffic demand	LSTM predicts the future traffic with higher accuracy
[117]	Joint user-slice pairing and association	Slice association and pairing algorithms for UEs	Higher system throughput and lowest latency as compared to benchmark algorithm	Suitable approach for massive traffic generation in runtime

matching diversity, CSI for each user (allocating resources) heuristic algorithm, etc. H-OMA and H-NOMA was found to be effective for scheduling resources in a mixed traffic scenario. Majority of the work discussed here are surrounded towards simulation environment where actual implementation of these work requires suitable planning and optimization of virtual resources for satisfying the QoS requirements. The hardware infrastructure for the network slicing also requires advance hardware infrastructure for real life deployments. Very few works were available based on nature-inspired algorithms for channel allocation and effective network slicing scheduling for virtual resource management.

3.4. Machine learning-based approaches

Machine learning is an effective approach to bring scalability and predictability to the coexistence mechanism between eMBB and URLLC. This section covers an extensive literature review based on machine learning approaches for the coexistence mechanism between eMBB and URLLC.

Tang et al. [120] formulated the uplink and downlink resource allocation problem using Time Division Duplexing (TDD) which is a full-duplex technology for the provisioning of the resources. They proposed a DRL-based algorithm for resource allocation in TDD environment. The algorithm adaptively allocates resources for the uplink and downlink transmission in a high mobility 5G heterogeneous network environment. The simulation result demonstrated a concise improvement in the throughput and packet drop ratio as compared to conventional Q-learning algorithms. The model can be further analyzed for reducing the computation capacity of the DRL-based resource allocation scheme. The proposed approach provides dynamic resource allocation while adapting TDD configuration with low overhead. The model was tested in low computational load and mobility. Furthermore, it needs to improve high mobility, low computational complexity and overhead for gaining more realistic picture.

Ali et al. [121] proposed a deep learning based resource allocation scheme to train the data set for resource optimization. They proposed solutions to the joint resource allocation problem and Radio Resource Head (RRH) association problems with a multi-tier C-RAN environment. An efficient sub-channel assignment, power allocation, and RRH association technique were used to generate the training dataset for the deep neural network (DNN) model with the iterative approach. The simulation results demonstrated that the prediction accuracy goes high as the number of samples and hidden layer increases, but the technique may initiate noisy learning features in several circumstances. The approach has good impact on the RRH association and power allocation in multi-RAT environments but introduces noisy learning feature when number of hidden layer reaches the threshold limit. It is good approach for an offline testing but in case of online test, it require adequate training on large network links and varying mobility and channel conditions.

Mikaeil et al. [122] focused on the uplink resource allocation strategy in a mobile C-RAN architecture. They formed a resource allocation as a NP-hard optimization problem. The proposed solution to the optimization problem is derived using the deep reinforcement learning approach. The performance of the algorithm was compared with two baseline algorithms viz. Genetic Algorithm and Tabu search. As compared to baseline algorithms they observed a significant improvement in the system throughput, fast convergence, and lower scheduling latency. The proposed algorithm also find out the adaptive allocation of transport block based on the actual radio block capacity of the Fronthaul link that significantly increases fronthaul bandwidth efficiency and decreased end-to-end uplink latency. Reinforcement learning is one of the most appropriate approaches for the resource allocation using C-RAN that provides lower delay, higher throughput, and good fairness index. However, the proposed approach only considered the uplink scenario, but for downlink scenario different sets of requirements must be fulfilled for effective resource allocation.

Designing a controller is a difficult task and required a lot of constraints related to the technology and policies, therefore in Ref. [123] authors presented a data-driven framework for a 5G network controller for link allocation. Experimental evaluations indicated that the evolved scheduler increased at a 150 % higher rate as compared to a single user, hence satisfying QoS requirements for the users. This approach shows the coexistence of Wi-Fi cell with LTE cells. It has higher latency as compared to other scheduler and for realistic environments. However, the proposed approach may be enhanced by considering dynamic parameters to support traffic and link variations.

Resource allocation in a network slicing environment is a challenging task due to the diverse traffic service requirements, different CSI parameters, and the mobility of the users. In this regard authors in Ref. [124] proposed a deep learning approach combined with reinforcement learning for resource allocation. By using an online time scheduling with inaccurate prediction and unexpected network they presented deep learning and reinforcement learning approaches to be responsible for large time scale resource allocation and small-time scale resource allocation respectively. They proposed a “actor and critic” algorithm for the resource scheduling on a small-time scale. Here the actor provides resource scheduling through policy network and critic uses time division error to make the decision more appropriate. For efficient resource allocation, the LSTM was utilized as a DL approach for the traffic prediction in the larger time scale. A significant performance improvement was notified in the proposed algorithm as compared to the other benchmark algorithms. The combination of deep learning and reinforcement learning is one of the emerging areas of research for effective resource allocation in coexistence mechanism. Practical implementation of these algorithm is quite challenging as it requires training on large amount of the datasets, also they should consider the complexity of system during real-life deployments. However, the prediction of low delay traffic help solving the coexistence problem of eMBB and URLLC traffic to some extent.

QoS is one of the important parameters that is used to map user service experiences. On the other hand, the admission control policy is helpful in proper resource allocation in terms of mapping dynamic and scalable network slices for the type of services required. In this regard, Vincenzi et al. [125] proposed a policy-based admission control mechanism for intra-slice allocation with an adaptable timescale. The algorithm worked under a predefined admission control policy that utilized offline data on resource availability and traffic load. The algorithm was

trained with a neural network to find the optimal admission control strategy in the run time. Simulation results demonstrated that the proposed algorithm has better performance as compared to the other offline algorithm in terms of admission rate, complexity, and low congestion level. The authors demonstrated the feasibility of the reservation-based slicing mechanism with strict QoS service requirements. However, in case of higher congestion, the model needs an adaptive congestion policy for QoS service guarantee.

Alsenwi et al. [126] formulated a dynamic multiplexing problem of the resource allocation and scheduling of the resources for eMBB and URLLC users. Firstly, resource blocks were allocated to the eMBB users based on the channel state and previous data rate up to the current timing. RBs allocation problem was further modeled into a two-dimensional Hopfield Neural Network (HNN) and the energy function of 2D-HNN is investigated to solve the problem. The scheduling problem was modeled as a convex optimization problem and that was solved using convex problem solver. The numerical results demonstrated that the proposed approach achieved a higher data rate for eMBB users and 90 % fairness to the RB allocation.

The authors in Ref. [127] considered a dynamic resource scheduling problem on the basis of time slots for guaranteeing the latency constraints of URLLC users and a higher achievable data rate for eMBB users. In addition, they considered the scheduling of the punctured eMBB users as a bandit problem. They model a priority selection strategy for scheduling the eMBB transmission drop while satisfying the URLLC requirements. The simulation results demonstrated an improvement in cumulative reward when they prioritized the scheduling of punctured eMBB resources. The model provides effective guidelines for victim eMBB as a bandit problem. The model was utilized only in downlink as it requires extensive research on uplink or mixed slot scenario. Victim eMBB on the other hand schedules traffic in next time slot that requires predictions about URLLC traffic in the resource grid. In case of unavailability of prediction as well as availability of URLLC traffic, the victim eMBB suffers another loss.

Li et al. [128] discussed on improving the limited transmission resource utilization according to the need of the users. In this regard, they formulated a reward function approach for different resource allocation policies. The arrival state was the Markov state, and the optimization problem was solved by introducing a Q-learning algorithm. The proposed Deep Q-Network (DQN) based algorithm achieved better performance in terms of intelligent resource scheduling, a trade-off between eMBB service queue length and URLLC service reliability. DQN can be important guideline for the operator but this scheme has serious implications such as complexity, exploration and exploitation to trade off problem, and training require for more adaptive data set based on the real network.

On the other hand, Filali et al. [129] proposed a two-level RB allocation to the users. To meet the QoS requirement satisfaction in term of data rate and delay, in the first level the larger time scale SDN controller allocates several RBs from the RBs resource pool as per the requirement of the gNB, whereas, in the second level the gNB allocates pre-allocated resources to the associated users. This mechanism reduced the gap between gNB and SDN controllers for the effective allocation of resources. In the first level, a single-agent RL-based algorithm was proposed for the partition of RBs among gNBs; whereas in the second level a multi-objective deep Q learning (DQL) based approach was proposed to share the RBs among gNBs. The proposed scheme shown an improvement in the data rate and latency of eMBB and URLLC in comparison with the benchmarked algorithms. The proposed approach has significantly reduced the complexity that was driven by generating an optimization problem. However, experimental study was concerned to low number of users, low mobility, and minimum packets load. Table 6 presents the key findings of several machine learning based approaches.

Table 6
Machine learning based approaches.

Reference	Problem Formulation	Techniques	Key Finding	Remarks
[49]	eMBB resource allocation and URLLC scheduling	Decomposition and relaxation-based resource allocation (DRPA) and combination of DRPA-PGACL algorithm	Reliability depends on the value of R_{min}. eMBB reliability decreased as R_{min} increased. It is also observed that increased URLLC traffic decreased the reliability of eMBB.	Increased average URLLC load decreases the eMBB sum rate.
[130]	To find the best modulation code scheme (MCS) so that loss is minimized	A DEMUX is designed to solve the preemption problem with the use of deep reinforcement learning. Learning of proposed method inspired from DDPG (Deep Deterministic Policy Gradient).	DEMUX provides better performance as URLLC load increases as compared to RP and FFP algorithms.	For 10 eMBB users, DEMUX obtained 76 % performance gain over FFP under the CDL model. For TDL it was found to be 81 %.
[131]	To update a dynamic Transmission time Interval (TTI) that will update according to the user requirements.	A flexible TTI strategy that satisfies the coexistence of eMBB and URLLC traffic service requirements. They implemented a Random Forest-based ensemble TTI decision algorithm for implementing flexible TTI.	The proposed algorithm outperforms as compared to fixed TTI, the delay performance of URLLC service improved by 45.64 %, and Packet loss Rate (PLR) increased by 59.17 % while guaranteeing the eMBB requirements.	Flexible TTI is not suitable for eMBB traffic, which has a large packet size, it is recommended to use fixed TTI (1 ms) for such eMBB services
[132]	Minimizing the impact of eMBB traffic symbol rate while considering URLLC traffic reliability	They proposed an algorithm that focused on the similarity region of the eMBB-URLLC symbol that led to less error rate instead of random selection	eMBB traffic symbol performance is measured in percentage of loss of eMBB symbols for the Enhanced Similarity region mapper (ESRM) are 18 % and 44 % compared to 59 % and 93 % for the URLLC mapper for BPSK-4QAM and BPSK-16QAM	The performance of eMBB strictly depends on the SNR and similarity region. If the SNR is increases the eMBB SER dominated by puncturing error.
[133]	Accurate slice management, Load Balancing, Slice failure, Allocate alternate slices during failure	A hybrid deep learning model that using CNN and LSTM.	The proposed model achieved an accuracy of 95.17 %.	The model provided no connection loss and optimum load balancing while considering new and outgoing routing request

(*continued on next page*)

Table 6 (*continued*)

Reference	Problem Formulation	Techniques	Key Finding	Remarks
[134]	Mobility of the user affects beam management and resource allocation	Deep Q-Learning with DBSCAN (DQLD) for effectively joint allocation of the resource with the help of online clustering mechanism	Better performance in terms of reliability, latency, rate of URLLC users and eMBB users.	The same architecture scenario can be used to train the model and find out the exact accuracy level
[135]	Resource allocation and scheduling problem for URLLC puncturing and eMBB transmission	Deep supervised learning	Higher accuracy of low complexity prediction for efficient resource allocation by adjusting the model parameter	Accuracy can be enhanced by adjusting the model paramters
[136]	Admission control problem	A novel framework based on machine learning using a 5G network with network slice management	The proposed technique outperforms in terms of prediction accuracy, resource smoothening, network throughput, and system utilization with the baseline approach	CPU utilization increases by enabling the resource manager scheme
[120]	TDD for uplink and downlink resource allocation	Deep Reinforcement learning for dynamic uplink and downlink resource allocation	System throughput increased and packet loss rate decreased	Computation overhead is more
[126]	RBs allocation problem as convex optimization problem	Modified 2D-HNN	90 % fairness increased among eMBB users	Higher achievable data rate of eMBB users was achieved in constraints to fairness
[137]	Resource allocation optimization problem	A multi-agent deep reinforcement learning	Enhanced URLLC reliability and QoS requirements for eMBB traffic	QoS satisfaction for eMBB and URLLC is a challenging task
[138]	Integer non-convex optimization	A multi-branch agent using DRL based on Branching Dueling Q-networks	The proposed DRL-based scheme provides higher reliability and increased service provisioning to URLLC	The mixed Numerology scheme increased performance due to inter-numerology interference mitigation
[139]	Multi-time scale problem	Hierarchical deep learning framework	Higher aggregate throughput and higher service level agreement	DRL algorithm is based on actor-critic method

3.4.1. Discussion

Based on the research literature, it can be stated that machine learning and deep learning approaches can significantly solve the optimization problems of radio resource allocation in eMBB and URLLC traffic scenarios. Most of the researchers worked on various techniques that were related to TDD, Markov decision process, deep reinforcement learning, LSTM, etc. to satisfy user's service requirements. Deep learning and reinforcement learning were found to be suitable to solve resource allocation problem and channel assignment among users. It is observed that the deep learning-based approaches were able to solve the large time scale (eMBB users) resource allocation problems, whereas reinforcement learning was found to be suitable for small-time scale (URLLC) resource allocation problem. A proper admission control policy is also equally important for the management of the radio resource among users. Very few works were observed related to admission control policy in mixed traffic scenarios because advanced reservation of the resource would not work in the case of URLLC scheduling. It is evident that during eMBB traffic puncturing, a lot of eMBB users got affected because of high preference to URLLC traffic. The scheduling policy should maintain a list of those affected users so that they can be scheduled before scheduling of next eMBB traffic while applying the puncturing technique. It is also evident from literature that such DRL and RL based approaches have higher computational capacity, so the researcher needs to focus to reduce the computational cost of these frameworks. In addition, it also requires a long training on different sets of data and dynamic implementation environments. A proper training on these models can effectively schedule the resource on eMBB and URLLC coexistence traffic. Traffic prediction problem of URLLC using machine learning technique can be resolved here with the help of LSTM network. However, the realistic implementation of these technique requires efficient radio resource planning and optimization.

3.5. C-RAN-based approaches

Centralized/Cloud Radio Access Network (C-RAN)-based architecture separate the data plane and control plane of the 5G architecture. It also supports network function virtualization, cloud computing, and other advanced paradigms. This section focused on the works related to C-RAN architecture based on the coexistence mechanism between eMBB and URLLC.

C-RAN architecture have different network components due to multi-cell infrastructure and cloud infrastructure. Kassab et al. [140] discussed the coexistence of eMBB and URLLC in C-RAN architecture using OMA and NOMA to validate the system performance. Here, eMBB user operate on long codewords that are spread in time and frequency domain, whereas the URLLC traffic is random that is being decoded to the edge for satisfying low latency requirements. eMBB traffic is influenced to interference management capabilities and centralize decoding at cloud. OMA on the other hand face difficulty in the case of URLLC due to lower error handling capability. As compared with OMA, the NOMA technique provides better error handling capabilities due to SIC algorithm thereby providing significant gain in the performance of URLLC in terms of latency and reliability. However, the larger value of q (traffic generation probability) indicates that the puncturing technique have worst performance in NOMA system.

C-RAN architecture is responsible to centralize the baseband processing of multiple base stations in a centralize cloud. In this regards, Jankovi et al. [95] formulated computational and radio resource allocation problem in virtualize RAN architecture. The problem was based on eMMB, URLLC and mMTC traffic provisioning and resource demand characterization. Based on that concept, a joint radio and resource allocation algorithm was design by the authors which provide QoS aware guaranteed resource allocation. To effectively allocate radio resources for different services like eMBB, mMTC and URLLC, they utilize puncturing technique for immediate URLLC transmission and static slicing technique for mMTC and eMBB coexistence. URLLC puncturing technique also affects the mMTC performance because of mini slot utilization via the URLLC. If mMTC traffic require relaxed QoS requirements, this technique provides better performance for URLLC. It is observed that the greater number of URLLC slot occupancy help satisfying the URLLC performance demands. But in case of higher number of URLLC occupancy, the overall performance of eMBB degrades significantly.

On the other hand, A. Salman in Ref. [141] discussed the Queuing model based on mobile edge computing, cloud data center, and C-RAN architecture. Prioritization of the traffic is computed and placed on the C-RAN architecture as per the demands. The proposed model effectively utilizes the dynamic slicing technique to allocate resources in a C-

RAN environment. They devise the closed mathematical expressions for the derived metrics such as system throughput, CPU utilization time, average response time, and average waiting time and system drop rate.

In RAN slicing multiple logical networks can be built over a single RAN infrastructure. However, C-RAN architecture is unable to meet the stingiest latency demand of URLLC traffic while considering massive fronthaul capacity requirements for eMBB traffic. Ahsan et al. [142] proposed a novel cloud Fog RAN over wavelength division multiplexing architecture. In this architecture RAN layers are divided into three layers viz. RRH, Fog node and BBU. Time sensitive URLLC traffic is handled by the Fog layer which is near to the RRH. On the other hand, eMBB and mMTC traffic is handled by BBU hotels. In comparison with the existing two-layer architecture, the performance of the proposed architecture was found good. It is observed that the value of weight derived (i.e., α) that was set to a constant value in this study has significant impact on the overall performance. Further the optimization to the value of α leads to the significant performance improvements.

Liu et al. [143] focused on the energy efficiency using distributed MIMO system for slicing of eMBB and URLLC traffic. They not only adopted a non-orthogonal slicing mechanism for improving the spectral efficiency but also short packet transmission mechanism for adopting URLLC requirements. The proposed approach has shown significant improvements in the energy efficiency and resource optimization as compared to the baseline approaches. It is observed from the study that the minimum weight factor for eMBB slice greatly impact the performance.

Similarly Setayesh et al. [144] proposed a mixed integer nonlinear programming that not only considers the joint energy efficiency but also the resource allocation. They proposed an algorithm based on penalized successive convex optimization to find the suboptimal solution of the proposed problem. In terms of network throughput, the proposed algorithm provides nearly 30 % improvement as compared to base line algorithm. The proposed algorithm was tested on the low traffic load of eMBB and URLLC. In case of high traffic loads, the proposed approach may face several challenges that leads to filth in the overall performance.

On the other hand, Domenico et al. [145] discussed the computational resource allocation and network slice deployment in hybrid C-RAN architecture. The proposed architecture consists of seven macro cells in which RRH provide access to eMBB, URLLC and mMTC slices. Each type of services has a tight latency constraint requirement. However, the VNF does not have optimal deployment of services.

Radio resource management scheme is one of the important concerns that needs to be considered while scheduling of the resources. In this regard, Kooshki et al. [146] presented an approach that combines both time domain and frequency domain schedulers for the coexistence of eMBB and URLLC. The novel technique shown nearly 29 % improvement in URLLC latency and 90 % in SNR of eMBB throughput. The proposed algorithm depicted an optimization in the resource allocation, but simulation was tested on fixed environments. For more lively environments, dynamic joint resource allocation technique is required that can adopt the dynamic behaviors of URLLC slice in C-RAN environments.

The realistic implementation of C-RAN requires effective planning and joint radio resource allocation with power efficiency as the main concern. A very few works were found to be focused on smart grid environment. Belaid et al. [147] designed a traffic framework for smart grid link scheduling and traffic routing in 5G integrated access and backhaul (IAB) network. The proposed approach provides significant improvement in energy efficiency, flow acceptance and achieved good network throughput as compared to baseline approaches. The traffic flow utilized here takes into account the coexistence between URLLC and eMBB. The proposed scheme utilized the puncturing approach instead of superposition technique because superposition technique creates unmanageable interference that is a big challenge for URLLC. However, the puncturing technique provide higher eMBB losses if URLLC load is higher.

On the other hand, in Ref. [148] authors mainly focused on the prediction and control of mobile traffic flow for the support of URLLC services in terms of high reliability, low latency, and extremely useability. The URLLC traffic does not maintain a queue, hence needs to be immediately scheduled without waiting. If the prediction of URLLC traffic is available, then it provides an additional advantage in terms of effective resource allocation for the mixed scenario. To obtain the URLLC traffic prediction, an LSTM-based approach was deployed on both edge cloud and remote cloud. The proposed mobile traffic prediction successfully predicts URLLC traffic and extensive simulation shows that the proposed algorithm outperforms in terms of latency and packet loss rate.

Network slicing brings virtualization concept in C-RAN architecture. According to the needs of users, the virtual slice can be scheduled by the gNB to achieve desired QoS requirements. In this regard, Boutiba et al. [149] proposed a new framework namely new radio flexibility (NR-flex) which indicate challenges faced by the network slicing in C-RAN environments. They utilize the concept of bandwidth part and dynamic numerology to meets the QoS demands for the end users. They define a multiplexing approach for the slicing where one bandwidth part is active at a time for a given user. As per 3GPP, the life cycle of slice is having four different phases viz. preparation, activation, run time, and decommissioning phase. Once these steps are completed the gNB allocate PRBs to these life cycle phases considering the result of preprocessing and multiplexing. The architecture brings flexibility for not only bandwidth part, but also the scheduler and preprocessor for resource allocation. However, for real time deployments it requires a critical examination to the detailed parameter of this architecture. In addition, the testing in dynamic environments is also essential.

On the other hand, Chen et al. [150] focused on twin GAN based DRL scheme for joint allocation of computational and bandwidth resources. The approach was based on two GAN based deep reinforcement learning models which jointly optimize and increase the spectrum efficiency thereby reducing the computational cost of computing resources. The performance metrices such as total delay, spectral efficiency, and computational cost reduction were improved by 10.2 %, 15.7 % and 12.8 % respectively. Distribution of all three types of services viz. eMBB, URLLC, and mMTC was utilized with different scenarios that considers varying loads of eMBB, URLLC and mMTC for effective resource allocation. By combining all three use cases, the proposed scheme becomes quite complex that utilize edge to cloud infrastructure for scheduling the resources. In real-life deployments it requires careful planning and optimization of network components and resources. Table 7 presents a summary of few C-RAN architecture-based coexistence mechanism between eMBB and URLLC traffics.

3.5.1. Discussion

The C-RAN architecture provides substantial benefits in terms of cost efficiency, network performance, scalability, flexibility, energy efficiency, and virtualization support. These benefits make C-RAN an appealing option for 5G NR deployments, particularly in crowded urban regions and situations requiring high capacity and performance. Coexistence of eMBB and URLLC can be efficiently addressed by exploiting C-RAN's centralized processing and resource management capabilities. Resource partitioning, dynamic resource allocation, network slicing, enhanced interference control, and QoS differentiation are crucial methods within the C-RAN architecture for ensuring efficient coexistence and meeting the specific requirements of both eMBB and URLLC services. Several concepts were discussed in the literature regarding radio resource allocation and computational resources optimization in C-RAN environments. Some of the technique mentioned here are queuing theory model which reflect the advantage of MEC server for low latency communication, resource allocation based on successive convex optimization which state the importance of admission control strategy, en-

Table 7
C-RAN based approaches.

Ref	Problem Formulation	Techniques	Key Finding	Remarks
[95]	Downlink radio and computational resource allocation problem	Joint computational and radio resource allocation framework	Computational load distribution algorithm balances the load as per traffic requirements. Radio resource algorithm implements eMBB-mMTC slicing and URLLC puncturing phenomena.	Performance degrades while reducing the resource allocation interval experience
[141]	Queuing Theory	A queuing model is proposed to analyze the performance of different slicing schemes	An analytical model derives the expressions for evaluating the performance of the system.	MEC server is important for low latency communication
[143]	Optimized beamforming design and RRU selection	Non-orthogonal scheduling to enable coexistence of eMBB and URLLC and short packet transmission for URLLC	Maximize the energy efficiency and guaranteeing the QoS requirements for eMBB and URLLC slice	Factors influencing energy efficiency are number of antennas, block length and error decoding capabilities
[151]	Orthogonal and non-orthogonal allocation of the resources for coexistence traffic	Heterogeneous-NOMA for utilizing the resources in a Cloud environment	Increased eMBB sum-rate due to effective utilization of spectrum	URLLC interference management on eMBB signals
[152]	Multi connectivity of the users with more than one base station	Multi-connectivity framework for URLLC users	Sum network throughput increased; Outage probability decreased	Performance depends on SINR, distance to UEs, and path loss model
[153]	Energy cost minimization problem formulated for short packet transmission as a mixed integer non-linear programming (MINLP) problem	Convex optimization tools and time-saving algorithm	The proposed algorithm saved energy with different simulation environments as compared to the baseline algorithms	Resource allocation subproblem is using fixed offloading strategy
[154]	The resource allocation problem is the mixed-integer nonlinear program problem	A resource allocation algorithm was proposed, based on a penalized successive convex approximation	30 % enhancement in throughput	Additional admission control strategy required for further improvement in the performance
[146]	eMBB and URLLC coexistence with QoS Satisfaction for each user	Enhanced radio resource management scheme in cell-less architecture	Average 90 % SINR level improvement over allocated RBs to schedule eMBB users	Need to study the impact of transmission time for scheduling resources in coexistence environments
[155]	eMBB and URLLC coexistence using OMA and NOMA	Uplink C-RAN theoretical model in fronthaul signaling	NOMA attained higher eMBB transmission rate in comparison to OMA and maintains a guaranteeing QoS service achievement for URLLC	The numerical results were demonstrated on the limited parameters
[156]	Minimization of energy consumption and maximization of system utility performance	Multifactor deep-*Q*-network resource allocation (mDQNR) framework for UEs and MEC server	Maximized system utility performance while minimizing energy efficiency	Framework supports AI based cross layer resource allocation

hanced cell less radio architecture which denote the importance of the transmission time in scheduling of the resources and most importantly energy efficiency which optimize the joint energy and resource allocation problem. Here again orthogonal and non-orthogonal resource allocation scheme have added advantage while utilizing C-RAN environment, in which NOMA successive management of interference from eMBB and URLLC coexistence. The use of MEC in C-RAN based approach is beneficial for URLLC traffic because these servers are near to the end user and eMBB traffic can be accommodate by the cloud server in which latency is not a constraint. However, MEC is unable to identify the traffic because operators take decisions to schedule the traffic on that node. Based on a variety of parameters such as network congestion, latency requirements, application characteristics, and available resources, the operator can decide whether to arrange traffic processing and computing jobs to either MEC or the cloud server. MEC (at the edge) is often utilized for low-latency applications that require real-time processing or local data offloading, whereas the Cloud server (centralized) provides greater computational capacity and scalability for more resource-intensive operations. But here effective offloading strategy requires predicting and scheduling of URLLC traffic at the mobile edge node.

3.6. Summary of other survey articles based on 5G NR

To the best of our knowledge, very few survey/review articles were available on 5G NR resource allocation schemes. In addition, we did not find any review article on coexistence mechanisms of eMBB and URLLC that is focused on QoS provisioning and optimization of resource allocations. This section highlights a summary of several survey articles focused on presenting the key contributions to 5G NR. 5G provides flexible numerology support, number of KPI, and QoS parameters for future generation mobile networks. In this regard, Dogra et al. [157] presented a summary of 93 research articles of ranging from 2009 to 2020. They provided a brief overview of 5G NR, its key performance indicators, and issues related to the adaption of future generation technology. In addition to 5G, they also emphasized various technologies and applications of the sixth-generation mobile network. Fog computing and MEC are important paradigms in 5G NR, therefore the authors in Ref. [158] presented a comprehensive survey on resource scheduling algorithms for fog and MEC technologies. They also focused on optimization metrics, evaluation tools in Fog computing, and Internet of Everything (IoE) environment. Based on the survey of the resource scheduling algorithms they mentioned some of the key issues and challenges faced by the researchers for adopting the technologies.

5G technology was deployed by many countries to fulfill the needs of every operator, but the coexistence technique is still one of the major

issues that was ignored by the service provider to adopt the 5G standard. Therefore Mamadou et al. [159] presented an extensive survey on the wireless network coexistence mechanism, especially on resource sharing in 5G era. The survey covers the coexistence of existing protocols, techniques, and mechanisms for 5G mobile network standards that impact existing 5G network infrastructure. They also discussed the open research issues and challenges for the future wireless resource allocation strategy. Table 8 presents existing survey with the major contribution of the different authors in their work.

URLLC and eMBB coexistence mechanism in 5G NR is the major research challenge because of their various trade-off between the performance metrics, therefore Khan et al. [160] presented a comprehensive survey on eMBB and URLLC coexistence mechanism for Industrial IoT (IIoT) environments. They discussed the importance of various key technologies that can make eMBB and URLLC services more diverse and investigated the trade-off between eMBB and URLLC coexistence mechanism. They also provided future directions to the researchers to optimize reliability and throughput metrics in IIoT.

Latency is one the main performance parameters in 5G NR. In this regard, Parvez et al. therefore [161] presented a comprehensive survey on the low latency toward 5G core network and solutions. They highlighted that to achieve a low latency communication, there is a need to change in network architecture including core, catching and RAN. On the other hand, Ali et al. [162] presented a survey based on tactile internet to achieve the low latency communication based on federated reinforcement learning FRL and ML techniques. Tactile internet is more advanced than the IoT environment which demands low latency. In addition, authors also identified future use cases for the tactile internet. Yin et al. [163] also presented a federated learning based scheme for power allocation and distributed spectrum Another comprehensive survey on tactile internet was presented by Sharma et al. [164]. They categorized the work on haptic communication, wireless AR/VR, and autonomous intelligent and cooperative mobility solutions. Several important issues such as MAC layer designing, and security and privacy aspects were also highlighted by authors.

It is evident that the coexistence mechanism between different use cases is a challenging task. In this regard, Pokhrel et al. [165] presented a survey on the mMTC and URLLC services. They identified the key challenges for implementing state-of-the-art technologies and their respective solutions. On the other hand, It is evident that time division duplexing strategy provides a dynamic resource allocation for uplink and downlink traffic. Resource management consists of managing user allocation, bandwidth, the transmission power of the antenna, and modulation scheme. In this regard Samidi et al. [166] presented a survey on various resource allocation strategies in a time division duplex environment. They categorized three categories for resource management viz. resource allocation, interference management, and energy efficiency. The survey included potential contributions and challenges faced by resource allocation strategies. In the other hand, NOMA technology significantly increases spectral efficiency and allocate resources effectively to the user. A comprehensive survey on multiple antenna techniques in NOMA is presented by Tian et al. in Ref. [167]. To solve the severe co-channel interference and high implementation complexity problem, NOMA utilizes multiple antenna techniques. The survey was conducted on multiple antenna techniques viz. two users, multi-users, and massive connectivity in a heterogeneous system using NOMA.

To work on the future generation of the mobile technologies needs to understand the existing technologies, therefore the Akhtar et al. [168] presented a survey on the radio resource management (RRM) approach toward 4G to 5G and beyond. They discussed various RRM schemes and highlighted several challenges to implementing the RRM strategy.

4. Commonly used simulation parameters for coexistence of eMBB and URLLC

This section illustrates various simulation parameters used by researchers in their work as mentioned in preceding sections. Table 9 presents simulation parameters that are suitable for eMBB and URLLC coexistence mechanisms.

5. Use case scenario of eMBB and URLLC coexistence

This section demonstrates the use case scenario of eMBB and URLLC coexistence concerning real life deployments. Plenty of use case scenarios exists that requires the coexistence of eMBB and URLLC in real time 5G network communications as depicted in Fig. 7.

These scenarios either focus on puncturing or superposition technique to schedule traffics in the resource grid. Some of the use-case scenarios are mentioned as follows.

- In the smart grid scenario, eMBB offers real-time monitoring and control of power grids, allowing vast amounts of data to be transmitted for grid optimization and energy management. For important control commands and defect detection, URLLC ensures dependable and low-latency communication [169].

Medical use cases allowing doctors to remotely access and analyses medical data by utilizing eMBB traffic that supports high-resolution medical imaging, telemedicine, remote patient monitoring, etc. URLLC communication enables dependable and low-latency connection for real.

- -time patient monitoring, surgical support, and emergency response systems [170].

Table 8
An overview of existing survey in 5G NR.

Ref.	Published Years	Sources	Years Range	Problem Formulation Classification	eMBB and URLLC Coexistence	Simulation Parameters	Issues and Challenges
[157]	2021	93	2009–2020				✓
[158]	2022	143	1992–2022			✓	✓
[159]	2020	84	2000–2022				✓
[160]	2022	215	2010–2022		✓		✓
[161]	2018	256	2006–2018				✓
[162]	2021	184	2002–2021				✓
[164]	2020	226	2010–2020				✓
[165]	2020	131	2001–2020				✓
[166]	2021	95	2012–2021				✓
[167]	2019	200	2007–2019				✓
[168]	2020	310	2000–2019				✓
This Work		203	2010–2022	✓	✓	✓	✓

Table 9
A summary of commonly used simulation parameters for coexistence of eMBB and URLLC.

Ref No	Frequency	SCS (kHz)	Cell Radius (Meter)	Time Slot Length (ms.)	TTI length (ms.)	Total System Bandwidth (MHz)	OFDM symbols in URLLC	URLLC packet Size (Bytes)	eMBB traffic model	Noise Floor/ Figure (dBm)
[99]	3.6 GHz 704 MHz	30	200	0.5	1	700	2	–	–	9 dB
[104]	–	15	100	1	1	48	2	10 kbits	–	−104
[86]	–	mixed	250	1	1	18	7	32	Fully buffered	–
[89]	2 GHz	15, 30, 60	200	0.25,0.5,1	1	720 kHz		–	–	7 dB
[90]	3.5 GHz	30		0.5	1	20	4	256 bits	CBR	−61
[91]	4 GHz	30	500	0.5	1	20	2	50	CBR	–
[92]	2 GHz	15	200	1	1	3	–	–	–	7 dB
[93]	2 GHz	15	100–360	1	1	20	2	100	Fully buffered	5/9 dB
[95]	–	15	–	1	1	–	2, 4, 7	10	Uniform	20 dB
[135]	–		500	–	1	–	–	20	–	18 dB
[130]	4 GHz	30	1000	0.5	1	20	2	50	Fully buffered	−91.9
[47]	10 MHz	120	200	0.125	1	50	2	32–200	Fully buffered	−114
[77]	–	15	250	1	0.5	20	2	32	Fully buffered	–
[74]	2 GHz	15	500	1	0.143	10	2	–	Fully buffered and CBR	–
[49]		15	300	1	1	20	2	32	Fully buffered	–

Fig. 7. Use Case scenario for eMBB and URLLC coexistence.

- The eMBB offers immersive Augmented reality and virtual reality (AR/VR) experiences by provide high-resolution video streaming, interactive content, multiplayer gaming, et. URLLC ensures low-latency communication for real-time interactions, reducing motion sickness and improving user experience [171].
- eMBB allows high-quality video monitoring, remote maintenance, and real-time data analytics for industrial automation systems. Whereas URLLC communication allows ultra-reliable, low-latency control and coordination of robots, machinery, and manufacturing processes.
- eMBB offers a variety of smart city applications, including intelligent traffic control, environmental monitoring, and networked infrastructure. Whereas, low-latency communication is enabled by URLLC for real-time traffic control, pedestrian safety, and emergency services [172].

- In precision agriculture eMBB offers agricultural field remote monitoring, crop analysis, and automated irrigation systems. Whereas, URLLC offers real-time control of drones, autonomous farm equipment, and sensor networks using low-latency communication [173].

6. Issues, challenges, and future research directions

The coexistence mechanism between eMBB and URLLC discussed in the literature review aroused various key issue and challenges for the researchers. This section discusses about the key issues, major challenges, and future direction in the coexistence mechanism between eMBB and URLLC services.

6.1. Issues and challenges

6.1.1. eMBB throughput and URLLC reliability

The puncturing methods applied to the coexistence mechanism between URLLC and eMBB traffic leads to the suspension of ongoing eMBB transmission and replaces the slots with a URLLC traffic, which results to a subsequent loss in the performance of eMBB users. One of the challenging task in 5G NR is to improve the eMBB throughput while puncturing the eMBB traffic [174]. If the number of URLLC traffic increases in bulk then it is very difficult to accommodate all URLLC traffic which leads to reliability loss of URLLC users [55,175]. Maintaining eMBB throughput and URLLC reliability is an open research challenge because both metrics depend on the traffic load. If the traffic load of URLLC goes high (more than 10 users simultaneously transferring the data), then it is difficult to maintain the URLLC reliability constraint [176]. The lack of channel information and dedicated bandwidth is also the major obstacle for URLLC in implementing the latency requirements [177].

6.1.2. Interference

Another key challenge in implementing the puncturing mechanism is related with the interference of URLLC and eMBB traffic. During puncturing of eMBB traffic, the power level of eMBB is set to zero for avoiding the interference. In this process, both the power and bandwidth allocation to the URLLC traffic incurs a significant delay that results in reduced reliability [178]. However, interference issue can be resolved using the NOMA technology [57] but it was found to be more effective for the uplink transmission [179,180]. The puncturing mechanism also provides a better eMBB throughput related to the URLLC latency constraints in the case of a trade-off factor applied to the selection of OMA and NOMA techniques [181]. In addition, several anti-interference measures [182] should be taken into account while designing a framework for eMBB and URLLC co-existence.

6.1.3. Error handling

There are several mechanisms exists in the research literature to handle the errors. HARQ is one of the successful protocol that provides a good error handling capability. For eMBB transmission, the BLER target for the CSI report should be less than10 % [175]. Handling errors in URLLC is another big challenge because most of the HARQ methods are not suitable for meeting URLLC requirements, especially for time-critical applications which generates uneven delay [183]. The BLER also depends on the choice of MCS used, so it needs to choose modulation and coding scheme for effectiveness in the case of URLLC transmission which has low retransmission urgency [184].

6.1.4. Frame design and packet size

Another important key issue for both the coexistence mechanism not only frame designing but also determining the size of the packet. eMBB packet needs to be scheduled in the given time slot and the URLLC packet to be scheduled in mini-slots [185]. 5G NR is using the non-square-shaped packet in the frequency domain. Here polar code is used for the control channel and a low-density parity check is utilized for the data channel [183]. The longer block length of the packets and frequent retransmission of the errored packet increased the latency of URLLC traffic.

6.1.5. Handover and user mobility

In 5G NR, one of the most critical parts is the handover with constraints to low latency and jitter when UEs nodes are mobile in nature and frequent handover phenomena happened [183]. To find out the BS during the handover process is another key issue because due to the dense infrastructure various BS cell coverage will be available to use [186]. Most of the URLLC nodes are mobile in nature and they move continuously at a high speed, so maintaining latency is a key issue during that time [187]. Different mixed scenario traffic has different QoS requirements in term of latency, reliability, and mobility. For e.g., a high-speed train have high mobility requirement in comparison to IoT devices that have low mobility but the massive infrastructure of the devices [188,189]. There are two main points regarding mobility management viz. location registration and handover management. Both of them are the key issues to satisfy the stringent latency and reliability requirements [190,191].

6.1.6. Resource allocation in NFV

SDN provides a network function virtualization services to the radio resources which is available in the central pool of the resource manager [192]. To meet the strict QoS requirements for different network slices and provisioning virtual resources to the slices is a very challenging task in 5G NR [188].

6.1.7. Inter-user and intra-user multiplexing

Inter-user and intra-user multiplexing is another key issue for simultaneous eMBB and URLLC traffic. There is a possibility that different UEs might prioritize the logical channel for different traffic scenarios before data transmission. One UE may have both eMBB and URLLC traffic in downlink and uplink transmission. Simultaneous handling of both downlink and uplink transmission with traffic characterization is very challenging task in 5G [193].

6.1.8. eMBB and URLLC scheduling

The coexistence of eMBB and URLLC traffic scenario requires an effective resource scheduling mechanism to satisfy the QoS requirements [194,195]. Plenty of scheduling algorithms were proposed in the research literature to satisfy the stringent URLLC latency, reliability, and eMBB throughput. But still there is a scope to enhance the eMBB throughput without disturbing URLLC reliability and latency [196]. There exist various key issues such as the proportional fair algorithm does not have a higher throughput for distance users and also facing resource blockage problem [197]. On the other hand, round robin scheduling algorithm have lack of channel state information. Best CQI algorithm always give preferences to the users who have good channel conditions; so such algorithms lacks in fairness and are not suitable for the traffic conditions that require strict latency [194]. It is evident that the fairness is inversely proportional to the throughput. If the higher fairness is deployed automatically the throughput is reduced and if users want a higher throughput, the fairness among users is reduced [44]. However, authors in Ref. [84] tried to resolve this issue by proposing a enhance joint scheduling algorithm that utilizes two scheduling approaches viz. the best CQI highest deviation and best CQI lowest second. Channel-aware scheduling algorithm utilizes a heuristic algorithm that is best suited for the coexistence of eMBB and URLLC scheduling [89].

6.1.9. Mixed numerology interference

5G utilizes eMBB, URLLC, and mMTC for the mixed traffic scenarios. As per 3GPP standards [16,17], the mixed numerology is defined in the 5G NR architecture [198]. It is apparent that the eMBB packet size is large enough that needs to be placed in the common slot of 1 ms (15 kHz SCS), whereas the URLLC can be placed in the mini slot [165]. Due to the mixed numerology this hybrid configuration can face serious interferences because there might be a chance that a single user can use both services at a same time. Multiplexing of the heterogeneous subcarrier generates inter numerology interference, which reduced the service provisioning of the eMBB and URLLC requirement [199]. It is needed to design an efficient slicing enforcement algorithm to prevent interference between different numerology.

6.1.10. Dynamic resource allocation

Dynamic resource allocation for the coexistence mechanism of eMBB and URLLC is another key challenge [200]. Most of the work was found to be focused in formulating the joint resource allocation problem as an optimization problem that can be solved using heuristic algorithms due to their low complexity and less execution time [112,154].

6.1.11. Framework for combing different technologies

To design an effective framework that supports several key approaches such as beamforming, NOMA, OMA, edge and fog computing in C-RAN architecture, distributed machine learning techniques, flexible TDD configurations, grant free access, network slicing, QoS requirements, and delay budget reporting is one of the major challenge in 5G [201].

6.1.12. Power allocation and priority slicing

A network that does not provide priority to the traffic and treats every traffic with equal hallmarks can significantly reduce the potentiality of the 5G network. URLLC has no queue and needs immediate scheduling due to the latency constraint, therefore, it is very challenging task to allocate adequate power level by gNB for such traffics [109,202]. The gNB needs to prioritize the traffic according to the information about UEs availability [98]. With the help of queuing mechanism, the gNB can easily prioritize the mMTC and eMBB traffic, but such type of queuing mechanism does not support URLLC traffic. To prioritize the URLLC users with minimum E2E delay requirements is one of the key challenges of 5G NR [203].

6.2. Future research directions

The eMBB and URLLC coexistence mechanism is a complex approach that require dynamic multiplexing schemes. These schemes should be flexible to accommodate sporadic URLLC traffic. In the preceding section, various key issues and challenges were highlighted that can be incorporated to enhance the performance in the coexistence mechanism. As mentioned in 3GPP standard, the puncturing technique is found to be an effective preemption approach for the resource allocation, but it lacks in to achieving higher eMBB throughput while maintaining URLLC latency and reliability constraints. To obtain better results, it is recommended to use a lower MCS level for eMBB transmission. In short, researchers may try to identify the best MCS approach so that the eMBB loss is minimized. In contrast, to devise an optimal QoS provisioning-based approach for improvement in the coexistence mechanism between eMBB and URLLC is also suggested. It is observed that the use of various machine learning and deep learning-based approaches can predict the URLLC traffic. However, with the help of these techniques a significant gain in the performance of the eMBB throughput was observed. But due to the slow convergence rate, deep reinforcement learning techniques were not able maintain the QoS provisioning. On the other hand, LSTM worked with a fine granularity level for the prediction of the URLLC traffic. A subsequent performance enhancement may be observed by combing the LSTM with Bayesian network. In addition to that, machine learning model are trained to specific set of data but 5G network is highly dynamic in nature due varying radio conditions, user density, traffic patterns. Self-adaption of these machine learning model requires higher computational complexity but unable to maintain the required latency. It is also difficult to maintain higher amount of data to train deep learning models. It is evident that URLLC devices have low latency requirements, but the incorporation of machine learning models and their subsequent learning mechanisms incurs additional computational cost thereby significantly delaying the overall performance. In future, the slice management and orchestration to ensures the service level agreement can be considered as open research issue.

7. Conclusion

The eMBB and URLLC traffic scenarios are suitable to most of the application areas such as industrial IoT, healthcare, autonomous vehicles, smart cities, etc. Effectively deploying the 5G use case scenario, services, and application support is found to be challenging for the service providers. In this paper, the coexistence mechanisms between the eMBB and URLLC services was investigated on five major classes viz. multiplexing-, QoS-, Machine learning-, Network slicing-, and C-RAN architecture-based approaches. As per 3GPP standardization, puncturing and superposition technique provide effective resource utilization and enable dynamic coexistence between eMBB and URLLC. Major issues were eMBB throughput loss and lack of URLLC reliability requirements. Whenever the traffic density increases, the eMBB loss increases to satisfy the URLLC users. To satisfy the eMBB and URLLC coexistence requirements, several enhancements (i.e., random puncturing, online gradient scheduling, grant based resource allocation grant free resource allocation, lower MCS value selection, joint scheduling algorithm etc) were proposed in state-of-the-art. Researchers also potentially contributed to various approaches based on NOMA, millimeter waves, RAN slicing, and mobile cloud and mobile edge computing. We highlighted various key issues and challenges that needs to be considered while deploying eMBB and URLLC coexistence. It was observed that a very few works were contributed to the field of QoS provisioning based coexistence mechanism. It can be stated that that there is enormous scope to improve the QoS parameters such as GBR, Non-GBR, and delay critical GBR for coexistence mechanism between eMBB and URLLC.

Disclosure

The authors declare that this work does not involve any survey or human participants or animals in any capacity.

Funding statement

The authors received no funding for this manuscript.

Conflicts of interest

The authors declare that they have no conflicts of interest to report regarding this study.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

All the data are shared in the main manuscript.

References

[1] D. Ku, H.W. Lee, S.C. Lee, S. Lee, 5G commercialization and trials in korea, Commun. ACM 63 (4) (2020) 82–85, https://doi.org/10.1145/3378430.

[2] D. Littman, P. Wilson, C. Wigginton, B. Haan, J. Fritz, 5G: the Chance to Lead for a Decade, Deloitte, 2018 [Online]. Available: https://www2.deloitte.com/us/en/pages/consulting/articles/5G-deployment-for-us.html.

[3] M. Attaran, The impact of 5G on the evolution of intelligent automation and industry digitization, J. Ambient Intell. Hum. Comput. (2021) 0123456789, https://doi.org/10.1007/s12652-020-02521-x.

[4] ETSI, TS 138 101-1 - V15.2.0 - 5G; NR; User Equipment (UE) Radio Transmission and Reception; Part 1: Range 1 Standalone (3GPP TS 38.101-1 Version 15.2.0 Release 15), vol. 15.2.0 2018 pp. 0–244. [Online]. Available: https://portal.etsi.org/TB/ETSIDeliverableStatus.aspx

[5] Y. Chen, Y. Wang, H. Wu, Research on F-OFDM subband filter algorithm for three major 5G application scenarios, ACM Int. Conf. Proceeding Ser. (2020), https://doi.org/10.1145/3448823.3448858.

[6] S. Hamdoun, A. Rachedi, Y. Ghamri-Doudane, Graph-based radio resource sharing schemes for MTC in d2d-based 5G networks, Mobile Network. Appl. 25 (3) (2020) 1095–1113, https://doi.org/10.1007/s11036-020-01527-1.

[7] Nokia White Paper, 5G use cases and requirements, https://www.ramonmillan.com/documentos/bibliografia/5GUseCases_Nokia.pdf.

[8] D. Martín-Sacristán, M. Michał, S.E. El Ayoubi, M. Fallgren, P. Spapis, 5G PPP use cases and performance evaluation models, 5G Infrastruct. Public Priv. Partnersh. (2016) 1–39 [Online]. Available: http://www.5g-ppp.eu/%0Ahttps://5g-ppp.eu/wp-content/uploads/2014/02/5G-PPP-use-cases-and-performance-evaluation-modeling_v1.0.pdf.

[9] K. Hiran, K. Lakhwani, J. Wireko, H. Gianey, Internet of Things (IoT): Principles, Paradigms and Applications of IoT, 2020.

[10] D. Marabissi, et al., A real case of implementation of the future 5G city, Future Internet 11 (1) (2018) 1–16, https://doi.org/10.3390/fi11010004.

[11] 5G Americas White Paper, "© Copyright 2017 5G Americas- 5G Services and Use Cases Nov 2017 1,", 2017, pp. 1–52.

[12] A.K. Bairagi, M.S. Munir, M. Alsenwi, N.H. Tran, C.S. Hong, A matching based coexistence mechanism between EMBB and URLLC in 5G wireless networks, Proc. ACM Symp. Appl. Comput. (2019) 2377–2383 https://doi.org/10.1145/3297280.3297513, Part F1477.

[13] M. Kq, "3GPP TSG RAN WG1 Meeting # 87 Agenda Item : Source : Title : AccelerComm on the Hardware Implementation of Channel Decoders for Short Block Lengths Document for : Discussion," No. November, 1–9, 2016.

[14] 3GPP R1-1700374, Downlink Multiplexing of eMBB and URLLC Transmissions, 2017, pp. 1–12 no. January.

[15] Prisma, PRISMA Statement, 2022. https://www.prisma-statement.org//PRISMAStatement/PRISMAStatement. (Accessed 20 July 2022).

[16] 3GPP TR 21.915, Digital cellular telecommunications system (phase 2+) (GSM); universal mobile telecommunications system (UMTS); LTE; 5G; release description; release 15, 3rd Gener. Partnersh. Proj. (3GPP), Tech. Rep. (2019) 1–120 21.915 version 15.0.0, vol. 0.

[17] E. Dahlman, S. Parkvall, J. Sköld, Chapter 5 - NR overview, in: E. Dahlman, S. Parkvall, J. Sköld (Eds.), 5G NR (Second Edition), Second Edi, Academic Press, 2021, pp. 57–78, https://doi.org/10.1016/B978-0-12-822320-8.00005-2.

[18] C. Campolo, A. Molinaro, F. Romeo, A. Bazzi, A.O. Berthet, 5G NR V2X: on the impact of a flexible numerology on the autonomous sidelink mode, in: IEEE 5G World Forum, 5GWF 2019 - Conf. Proc, 2019, pp. 102–107, https://doi.org/10.1109/5GWF.2019.8911694.

[19] X. Lin, et al., 5G new radio: unveiling the essentials of the next generation wireless access technology, IEEE Commun. Stand. Mag. 3 (3) (2019) 30–37, https://doi.org/10.1109/MCOMSTD.001.1800036.

[20] A. Omri, M. Shaqfeh, A. Ali, H. Alnuweiri, Synchronization procedure in 5G NR systems, IEEE Access 7 (2019) 41286–41295, https://doi.org/10.1109/ACCESS.2019.2907970.

[21] A. Yazar, B. Peköz, H. Arslan, Fundamentals of Multi-Numerology 5G New Radio, 2018 [Online]. Available: http://arxiv.org/abs/1805.02842.

[22] 3GPP, Physical layer procedures for data, ETSI Tech. Specif. Release 15 (2018) 1–94 [Online]. Available: https://portal.etsi.org/TB/ETSIDeliverableStatus.aspx.

[23] ETSI 3GPP, 5G NR Physical Layer Procedures for Data (3GPP TS 38.214 Version 15.4.0 Release 15), France, 2019 [Online]. Available: https://portal.etsi.org/TB/ETSIDeliverableStatus.aspx.

[24] D. Panno, S. Riolo, An enhanced joint scheduling scheme for GBR and non-GBR services in 5G RAN, Wireless Network 26 (4) (2020) 3033–3052, https://doi.org/10.1007/s11276-020-02257-8.

[25] H. Huawei, DL Multiplexing between URLLC and eMBB, 2017 3GPP TSG R, no. February.

[26] T.K. Le, U. Salim, F. Kaltenberger, An overview of physical layer design for ultra-reliable low-latency communications in 3GPP releases 15, 16, and 17, IEEE Access 9 (2021) 433–444, https://doi.org/10.1109/ACCESS.2020.3046773.

[27] 3GPP, "5G; Study on New Radio (NR) Access Technology," Etsi Tr 138 912 V15.0.0, vol. 0, pp. 0–76 2018 [Online]. Available: http://www.etsi.org/standards-search

[28] J. Peisa, et al., 5G Evolution: 3GPP Releases 16 & 17 Overview, 2020 [Online]. Available: https://www.ericsson.com/en/reports-and-papers/ericsson-technology-review/articles/5g-nr-evolution.

[29] H. Huawei, February, DL URLLC Multiplexing Considerations, 3GPP, TSG R, 2017 [Online]. Available: https://www.3gpp.org/ftp/tsg_ran/WG1_RL1/TSGR1_87/Docs/R1-1611222.zip.

[30] Y. Cheng, F. Cheng, B. Deng, J. Li, C. Mei, ARQ algorithm optimization of radio link control layer in 5G system, ACM Int. Conf. Proceeding Ser. (2020) 135 https://doi.org/10.1145/3425329.3425337, –140.

[31] T. Specification, TS 123 501 - V16.6.0 - 5G; System Architecture for the 5G System (5GS) (3GPP TS 23.501 Version 16.6.0 Release 16, 0, 2021, p. 251 [Online]. Available: https://www.etsi.org/deliver/etsi_ts/123500_123599/123501/15.09.00_60/ts_123501v150900p.pdf.

[32] Y. Boujelben, Scalable and QoS-aware resource allocation to heterogeneous traffic flows in 5G, IEEE Internet Things J. 8 (20) (2021) 15568–15581, https://doi.org/10.1109/JIOT.2021.3074111.

[33] 5G NR QoS architecture, QoS attribute and QoS flow - techplayon, http://www.techplayon.com/5g-nr-qos-architecture-qos-attribute-and-qos-flow/. (Accessed 5 April 2021).

[34] T. Specification, TS 138 410 - V16.3.0 - 5G; NG-RAN; NG General Aspects and Principles, 0, 0 Release 16, 2020, pp. 1–18 3GPP TS 38.410 version 16.3.

[35] ETSI, - 5G; NR; NR and NG-RAN Overall description, TS 138 300 - V16.2.0 0 (2020) 11–17 Stage-2 (3GPP TS 38.300 version 16.2.0 Release 16).

[36] A. Kumar, Quality of service (QoS) in 5G network – ashok kumar, https://ashok100.in/2021/01/31/quality-of-service-qos-in-5g-network/.

[37] S.C. Tseng, Z.W. Liu, Y.C. Chou, C.W. Huang, Radio resource scheduling for 5G NR via deep deterministic policy gradient, in: 2019 IEEE Int. Conf. Commun. Work. ICC Work. 2019 - Proc., 2019, pp. 1–6, https://doi.org/10.1109/ICCW.2019.8757174.

[38] A. Vora, K.-D. Kang, Effective 5G wireless downlink scheduling and resource allocation in cyber-physical systems, Technologies 6 (4) (2018) 105, https://doi.org/10.3390/technologies6040105.

[39] S.N.K. Marwat, M. Shuaib, S. Ahmed, A. Hafeez, M. Tufail, Medium access-based scheduling scheme for cyber physical systems in 5G networks, Electron 9 (4) (2020), https://doi.org/10.3390/electronics9040639.

[40] D. Kesavan, E. Periyathambi, A. Chokkalingam, A proportional fair scheduling strategy using multiobjective gradient-based African buffalo optimization algorithm for effective resource allocation and interference minimization, Int. J. Commun. Syst. 35 (1) (2022) e5003, https://doi.org/10.1002/dac.5003.

[41] L. Li, W. Shao, X. Zhou, A flexible scheduling algorithm for the 5th-generation networks, Intell. Converg. Networks 2 (2) (2021) 101–107, https://doi.org/10.23919/icn.2020.0017.

[42] O. Al Taee, A. Al Janaby, Y. Abbosh, 5G Uplink Performance of Symbol-Based Schedulers with Network Slicing (2022), https://doi.org/10.4108/eai.7-9-2021.2314842.

[43] Q. Ai, P. Wang, F. Liu, Y. Wang, F. Yang, J. Xu, QoS-guaranteed cross-layer resource allocation algorithm for multiclass services in downlink LTE system, in: 2010 Int. Conf. Wirel. Commun. Signal Process. WCSP 2010, 2, 2010 https://doi.org/10.1109/WCSP.2010.5633846, 0–3.

[44] D. Panno, S. Riolo, A new joint scheduling scheme for GBR and non-GBR services in 5G RAN, in: 2018 IEEE 23rd International Workshop on Computer Aided Modeling and Design of Communication Links and Networks, CAMAD), 2018, pp. 1–6, https://doi.org/10.1109/CAMAD.2018.8514964.

[45] A. Karimi, K.I. Pedersen, N.H. Mahmood, G. Pocovi, P. Mogensen, Efficient low complexity packet scheduling algorithm for mixed URLLC and eMBB traffic in 5G, IEEE Veh. Technol. Conf. (2019) https://doi.org/10.1109/VTCSpring.2019.8746407, 2019-April.

[46] N. Patriciello, S. Lagen, L. Giupponi, B. Bojovic, The impact of NR scheduling timings on end-to-end delay for uplink traffic, in: 2019 IEEE Glob. Commun. Conf. GLOBECOM 2019 - Proc., 2019 https://doi.org/10.1109/GLOBECOM38437.2019.9013231, no. c.

[47] A.K. Bairagi, et al., Coexistence mechanism between eMBB and uRLLC in 5G wireless networks, IEEE Trans. Commun. 69 (3) (2021) 1736–1749, https://doi.org/10.1109/TCOMM.2020.3040307.

[48] R.-1612537 3GPP, TSG RAN WG1 Meeting #87, "Evaluation Results of Superposition for Multiplexing eMBB and URLLC,", 2016, pp. 2–6 November.

[49] M. Alsenwi, N.H. Tran, M. Bennis, S.R. Pandey, A.K. Bairagi, 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. 20 (7) (2021) 4585–4600, https://doi.org/10.1109/TWC.2021.3060514.

[50] T. 3GPP Specification and R. 16 Services, Group, "3gpp Tr 21.916," vol. 1, no. Release 16 2021

[51] M. Baker, "5G Releases 16 and 17 in 3GPP," No, Release, 2016.

[52] J. Peisa, NR Evolution – Realizing the Full Potential of 5G 5G – beyond Mobile Broadband, 2018.

[53] T.3G.P.P. Specification, TS 123 501 - V17.4.0 - 5G; System Architecture for the 5G System (5GS) (3GPP TS 23.501 Version 17.4.0 Release 17, 0, 2022.

[54] T. Specification, TS 138 300 - V17.0.0 - 5G; NR; NR and NG-RAN Overall Description, 0, 2022 Stage-2 (3GPP TS 38.300 version 17.0.0 Release 17).

[55] A. Anand, G. De Veciana, S. Shakkottai, Joint scheduling of URLLC and eMBB traffic in 5G wireless networks, IEEE/ACM Trans. Netw. 28 (2) (2020) 477–490, https://doi.org/10.1109/TNET.2020.2968373.

[56] M. Han, J. Lee, M. Rim, C.G. Kang, Dynamic bandwidth part allocation in 5G ultra reliable low latency communication for unmanned aerial vehicles with high data rate traffic, Sensors 21 (4) (2021) 1–13, https://doi.org/10.3390/s21041308.

[57] Q. Chen, J. Wang, H. Jiang, URLLC and eMBB Coexistence in MIMO Non-orthogonal Multiple Access Systems, 2016, pp. 1–13.

[58] A.K. Bairagi, S. Munir, S.F. Abedin, C.S. Hong, Coexistence of eMBB and uRLLC in 5G Wireless Networks, 2018, pp. 1377–1379.

[59] I.S. Gerasin, A.N. Krasilov, E.M. Khorov, Dynamic multiplexing of URLLC traffic and eMBB traffic in an uplink using nonorthogonal multiple access, J. Commun.

Technol. Electron. 65 (6) (2020) 750–755, https://doi.org/10.1134/S1064226920060108.
[60] E.N. Tominaga, H. Alves, R.D. Souza, J. Luiz Rebelatto, M. Latva-Aho, Non-orthogonal multiple access and network slicing: scalable coexistence of eMBB and URLLC, IEEE Veh. Technol. Conf. (2021) https://doi.org/10.1109/VTC2021-Spring51267.2021.9448942, 2021-April.
[61] W. Sui, X. Chen, S. Zhang, Z. Jiang, S. Xu, Energy-efficient resource allocation with flexible frame structure for hybrid eMBB and URLLC services, IEEE Trans. Green Commun. Netw. 5 (1) (2021) 72–83, https://doi.org/10.1109/TGCN.2020.3028202.
[62] X. Huang, D. Zhang, S. Tang, Q. Chen, J. Zhang, Fairness-based distributed resource allocation in two-tier heterogeneous networks, IEEE Access 7 (2019) 40000–40012, https://doi.org/10.1109/ACCESS.2019.2905038.
[63] A. Pratap, R. Misra, S.K. Das, Maximizing fairness for resource allocation in heterogeneous 5G networks, IEEE Trans. Mobile Comput. 20 (2) (2021) 603–619, https://doi.org/10.1109/TMC.2019.2948877.
[64] N. Zarin, A. Agarwal, Hybrid radio resource management for time-varying 5G heterogeneous wireless access network, IEEE Trans. Cogn. Commun. Netw. 7 (2) (2021) 594–608, https://doi.org/10.1109/TCCN.2021.3063132.
[65] Y. Li, Y. Zhao, J. Li, J. Zhang, X. Yu, J. Zhang, Side Channel attack-aware resource allocation for URLLC and eMBB slices in 5G RAN, IEEE Access 8 (2020) 2090–2099, https://doi.org/10.1109/ACCESS.2019.2962179.
[66] L. Miuccio, D. Panno, S. Riolo, A new contention-based PUSCH resource allocation in 5G NR for mMTC scenarios, IEEE Commun. Lett. 25 (3) (2021) 802–806, https://doi.org/10.1109/LCOMM.2020.3040504.
[67] W.R. Wu, T.H. Lin, Y.H. Lee, An indicator-free eMBB and URLLC multiplexed downlink system with correlation-based SFBC, ACM Int. Conf. Proceeding Ser. (2019) 105 https://doi.org/10.1145/3369555.3369560, –110.
[68] I.A. Pastushok, N.A. Boikov, N.A. Yankovskii, Bit stream multiplexing in 5G networks, 2020 Wave Electron. its Appl. Inf. Telecommun. Syst. WECONF (2020) 5–8 https://doi.org/10.1109/WECONF48837.2020.9131511, 2020.
[69] H. Chen, J. Wu, T. Shimomura, New reference signal design for URLLC and eMBB multiplexing in new radio wireless communications, IEEE Int. Symp. Pers. Indoor Mob. Radio Commun. PIMRC (2018) 1220–1225 https://doi.org/10.1109/PIMRC.2018.8580882, 2018-Septe.
[70] Y. Yang, K. Hiltunen, F. Chernogorov, On the performance of Co-existence between public eMBB and non-public URLLC networks, IEEE Veh. Technol. Conf. (2021) https://doi.org/10.1109/VTC2021-Spring51267.2021.9448841, 2021-April.
[71] E.C. Chukwu, U.S. Abdullahi, G. Koyunlu, J. Sanusi, G. Sani, I.A. Gangfada, Performance evaluation of multiplexed 5G-new radio network services of different usage scenarios, Proc. 5th Int. Conf. Commun. Electron. Syst. ICCES 2020 (2020) 335–342 https://doi.org/10.1109/ICCES48766.2020.09138021, Icces.
[72] K. Ganesan, T. Soni, S. Nunna, A.R. Ali, Poster: a TDM approach for latency reduction of ultra-reliable low-latency data in 5G, IEEE Veh. Netw. Conf. VNC 0 (2016) 4–5, https://doi.org/10.1109/VNC.2016.7835946.
[73] A.A. Esswie, K.I. Pedersen, Multi-user preemptive scheduling for critical low latency communications in 5G networks, Proc. - IEEE Symp. Comput. Commun. (2018) 136–141 https://doi.org/10.1109/ISCC.2018.8538471, 2018-June.
[74] A.A. Esswie, K.I. Pedersen, Opportunistic spatial preemptive scheduling for URLLC and eMBB coexistence in multi-user 5G networks, IEEE Access 6 (2018) 38451–38463, https://doi.org/10.1109/ACCESS.2018.2854292.
[75] K. Zhang, X. Xu, J. Zhang, B. Zhang, X. Tao, Y. Zhang, Dynamic multiconnectivity based joint scheduling of eMBB and uRLLC in 5G networks, IEEE Syst. J. 15 (1) (2021) 1333–1343, https://doi.org/10.1109/JSYST.2020.2977666.
[76] X. Song, M. Yuan, Performance analysis of one-way highway vehicular networks with dynamic multiplexing of eMBB and URLLC traffics, IEEE Access 7 (2019) 118020–118029, https://doi.org/10.1109/ACCESS.2019.2937470.
[77] P. Korrai, E. Lagunas, S.K. Sharma, S. Chatzinotas, A. Bandi, B. Ottersten, A RAN resource slicing mechanism for multiplexing of eMBB and URLLC services in OFDMA based 5G wireless networks, IEEE Access 8 (2020) 45674–45688, https://doi.org/10.1109/ACCESS.2020.2977773.
[78] W.U. Khan, X. Li, A. Ihsan, Z. Ali, B.M. Elhalawany, G.A.S. Sidhu, Energy efficiency maximization for beyond 5G NOMA-enabled heterogeneous networks, Peer-to-Peer Netw. Appl 14 (5) (2021) 3250–3264, https://doi.org/10.1007/s12083-021-01176-5.
[79] A. Slalmi, H. Chaibi, R. Saadane, A. Chehri, Call admission control optimization in 5G in downlink single-cell MISO system, Procedia Comput. Sci. 192 (2021) 2502–2511, https://doi.org/10.1016/j.procs.2021.09.019.
[80] M. Darabi, V. Jamali, L. Lampe, R. Schober, Hybrid puncturing and superposition scheme for joint scheduling of URLLC and eMBB traffic, IEEE Commun. Lett. 26 (5) (2022) 1081–1085, https://doi.org/10.1109/LCOMM.2022.3149170.
[81] Y. Prathyusha, T.L. Sheu, Coordinated resource allocations for eMBB and URLLC in 5G communication networks, IEEE Trans. Veh. Technol. 71 (8) (2022) 8717–8728, https://doi.org/10.1109/TVT.2022.3176018.
[82] H. Zarini, N. Gholipoor, M.R. Mili, Resource Management for Multiplexing eMBB and URLLC Services over RIS-Aided THz Communication, 2023, https://doi.org/10.1109/TCOMM.2023.3233988.
[83] D. Ivanova, E. Markova, D. Moltchanov, R. Pirmagomedov, Y. Koucheryavy, K. Samouylov, Performance of priority-based traffic coexistence strategies in 5G mmWave industrial deployments, IEEE Access 10 (2022) 9241–9256, https://doi.org/10.1109/ACCESS.2022.3143583.
[84] D. Panno, S. Riolo, An enhanced joint scheduling scheme for GBR and non-GBR services in 5G RAN, Wireless Network 26 (4) (2020) 3033–3052, https://doi.org/10.1007/s11276-020-02257-8.
[85] X. Zhang, J. Wang, H.V. Poor, Heterogeneous statistical-QoS driven resource allocation over mmWave massive-MIMO based 5G mobile wireless networks in the non-asymptotic regime, IEEE J. Sel. Area. Commun. 37 (12) (2019) 2727–2743, https://doi.org/10.1109/JSAC.2019.2947941.
[86] P.K. Korrai, E. Lagunas, A. Bandi, S.K. Sharma, S. Chatzinotas, Joint power and resource block allocation for mixed-numerology-based 5G downlink under imperfect CSI, IEEE Open J. Commun. Soc. 1 (July) (2020) 1583–1601, https://doi.org/10.1109/OJCOMS.2020.3029553.
[87] N. Zarin, A. Agarwal, QoS based joint radio resource allocation for multi-homing calls in heterogeneous wireless access network, in: MobiWac 2018 - Proc. 16th ACM Int. Symp. *Mobil. Manag. Wirel. Access*, 2018, pp. 37–42, https://doi.org/10.1145/3265863.3265877.
[88] M. Gharam, N. Boudriga, Game theoretical model for resource allocation in 5G hybrid HetNets, ACM Int. Conf. Proceeding Ser. (2019) https://doi.org/10.1145/3320326.3320354, Part F1481.
[89] L. Miuccio, D. Panno, P. Pisacane, S. Riolo, Channel-Aware and QoS-Aware Downlink Resource Allocation for Multi-Numerology Based 5G NR Systems, 2021, pp. 1–8, https://doi.org/10.1109/medcomnet52149.2021.9501268.
[90] A.A. Esswie, K.I. Pedersen, Analysis of outage latency and throughput performance in industrial factory 5G TDD deployments, IEEE Veh. Technol. Conf. (2021) https://doi.org/10.1109/VTC2021-Spring51267.2021.9448733, 2021-April.
[91] G. Pocovi, A.A. Esswie, K.I. Pedersen, Channel quality feedback enhancements for accurate URLLC link adaptation in 5G systems, IEEE Veh. Technol. Conf. (2020) https://doi.org/10.1109/VTC2020-Spring48590.2020.9128909, 2020-May.
[92] A. Akhtar, H. Arslan, Downlink resource allocation and packet scheduling in multi-numerology wireless systems, in: 2018 IEEE Wirel. Commun. Netw. Conf. Work. WCNCW 2018, 2018, pp. 362–367, https://doi.org/10.1109/WCNCW.2018.8369012.
[93] I. Gerasin, A. Krasilov, E. Khorov, Flexible multiplexing of grant-free URLLC and eMBB in uplink, IEEE Int. Symp. Pers. Indoor Mob. Radio Commun. PIMRC 978 (2020) https://doi.org/10.1109/PIMRC48278.2020.9217168, 2020-Augus.
[94] T.K. Le, U. Salim, F. Kaltenberger, Improving ultra-reliable low-latency communication in multiplexing with enhanced mobile broadband in grant-free resources, IEEE Int. Symp. Pers. Indoor Mob. Radio Commun. PIMRC (2019) 1–6 https://doi.org/10.1109/PIMRC.2019.8904368, 2019-Septe.
[95] J. Jankovic, Z. Ilic, A. Oracevic, S.M. Ahsan Kazmi, R. Hussain, Effects of differentiated 5G services on computational and radio resource allocation performance, IEEE Trans. Netw. Serv. Manag. 18 (2) (2021) 2226–2241, https://doi.org/10.1109/TNSM.2021.3060865.
[96] M. Darabi, L. Lampe, Multi objective resource allocation for joint eMBB and URLLC traffic with different QoS requirements, in: 2019 IEEE Globecom Work. GC Wkshps 2019 - Proc., 2019, pp. 1–6, https://doi.org/10.1109/GCWkshps45667.2019.9024527.
[97] M.L. Attiah, A.A.M. Isa, Z. Zakaria, M.K. Mohsen, M.K. Abdulhameed, A.M. Dinar, Joint QoE-based user association and efficient cell–carrier distribution for enabling fully hybrid spectrum sharing approach in 5G mmWave cellular networks, Wireless Network 25 (8) (2019) 5027–5043, https://doi.org/10.1007/s11276-019-02109-0.
[98] S.A. AlQahtani, A.S. Altamrah, Supporting QoS requirements provisions on 5G network slices using an efficient priority-based polling technique, Wireless Network 25 (7) (2019) 3825–3838, https://doi.org/10.1007/s11276-018-01917-0.
[99] I. Vila, J. Perez-Romero, O. Sallent, A. Umbert, Characterization of radio access network slicing scenarios with 5G QoS provisioning, IEEE Access 8 (2020) 51414–51430, https://doi.org/10.1109/ACCESS.2020.2980685.
[100] H. Althumali, M. Othman, N.K. Noordin, Z.M. Hanapi, Priority-based load-adaptive preamble separation random access for QoS-differentiated services in 5G networks, J. Netw. Comput. Appl. 203 (March) (2022), https://doi.org/10.1016/j.jnca.2022.103396.
[101] D. Shen, T. Zhang, J. Wang, Q. Deng, S. Han, X.S. Hu, QoS Guaranteed Resource Allocation for Coexisting eMBB and URLLC Traffic in 5G Industrial Networks, 2022, pp. 81–90, https://doi.org/10.1109/rtcsa55878.2022.00015.
[102] Y. Zhao, X. Chi, L. Qian, Y. Zhu, F. Hou, Resource allocation and slicing puncture in cellular networks with eMBB and URLLC terminals coexistence, IEEE Internet Things J. 9 (19) (2022) 18431–18444, https://doi.org/10.1109/JIOT.2022.3160647.
[103] T. De Cola, I. Bisio, QoS optimisation of eMBB services in converged 5G-satellite networks, IEEE Trans. Veh. Technol. 69 (10) (2020) 12098–12110, https://doi.org/10.1109/TVT.2020.3011963.
[104] F. Song, J. Li, C. Ma, Y. Zhang, L. Shi, D.N.K. Jayakody, Dynamic virtual resource allocation for 5g and beyond network slicing, IEEE Open J. Veh. Technol. 1 (2020) 215–226, https://doi.org/10.1109/OJVT.2020.2990072.
[105] L. Feng, Y. Zi, W. Li, F. Zhou, P. Yu, M. Kadoch, Dynamic resource allocation with RAN slicing and scheduling for uRLLC and eMBB hybrid services, IEEE Access 8 (2020) 34538–34551, https://doi.org/10.1109/ACCESS.2020.2974812.
[106] S.O. Oladejo, O.E. Falowo, Latency-aware dynamic resource allocation scheme for multi-tier 5G network: a network slicing-multitenancy scenario, IEEE Access 8 (2) (2020) 74834–74852, https://doi.org/10.1109/ACCESS.2020.2988710.
[107] F. Fossati, S. Moretti, P. Perny, S. Secci, Multi-resource allocation for network slicing, IEEE/ACM Trans. Netw. 28 (3) (2020) 1311–1324, https://doi.org/10.1109/TNET.2020.2979667.
[108] E.J. Dos Santos, R.D. Souza, J.L. Rebelatto, H. Alves, Network slicing for URLLC

and eMBB with max-matching diversity channel allocation, IEEE Commun. Lett. 24 (3) (2020) 658–661, https://doi.org/10.1109/LCOMM.2019.2959335.
[109] A. Hossain, N. Ansari, Priority-based downlink wireless resource provisioning for radio access network slicing, IEEE Trans. Veh. Technol. 70 (9) (2021) 9273–9281, https://doi.org/10.1109/TVT.2021.3095901.
[110] X. Chi, Y. Jing, H. Sun, B. Yu, A random compensation scheme for 5G slicing under statistical delay-QoS constraints, IEEE Access 8 (2020) 195197–195205, https://doi.org/10.1109/ACCESS.2020.3033321.
[111] W. Shi, et al., Two-level soft RAN slicing for customized services in 5G-and-beyond wireless communications, IEEE Trans. Ind. Inf. 3203 (c) (2021) 1–10, https://doi.org/10.1109/TII.2021.3083579.
[112] T. Ma, Y. Zhang, F. Wang, D. Wang, D. Guo, Slicing resource allocation for eMBB and URLLC in 5G RAN, Wireless Commun. Mobile Comput. 2020 (2020), https://doi.org/10.1155/2020/6290375.
[113] K. Xiong, S.S.R. Adolphe, G.O. Boateng, G. Liu, G. Sun, Dynamic resource provisioning and resource customization for mixed traffics in virtualized radio access network, IEEE Access 7 (2019) 115440–115453, https://doi.org/10.1109/ACCESS.2019.2935606.
[114] J. Zheng, P. Caballero, G. De Veciana, S.J. Baek, A. Banchs, Statistical multiplexing and traffic shaping games for network slicing, IEEE/ACM Trans. Netw. 26 (6) (2018) 2528–2541, https://doi.org/10.1109/TNET.2018.2870184.
[115] D. Wu, Z. Zhang, S. Wu, J. Yang, R. Wang, Biologically inspired resource allocation for network slices in 5G-enabled internet of things, IEEE Internet Things J. 6 (6) (2019) 9266–9279, https://doi.org/10.1109/JIOT.2018.2888543.
[116] F. Kavehmadavani, V.-D. Nguyen, T.X. Vu, S. Chatzinotas, Intelligent traffic steering in beyond 5G open RAN based on LSTM traffic prediction, 1–29. http://arxiv.org/abs/2210.08829, 2022.
[117] M.A. Riad, O. El-Ghandour, A.M. Abd El-Haleem, Joint user-slice pairing and association framework based on H-NOMA in RAN slicing, Sensors 22 (19) (2022) 1–25, https://doi.org/10.3390/s22197343.
[118] Y. Kim, H. Lim, Multi-agent reinforcement learning-based resource management for end-to-end network slicing, IEEE Access 9 (2021) 56178–56190, https://doi.org/10.1109/ACCESS.2021.3072435.
[119] J.P. de B. Gonçalves, H.C. de Resende, R. da S. Villaca, E. Municio, C.B. Both, J.M. Marquez-Barja, Distributed network slicing management using blockchains in E-health environments, Mobile Network. Appl. 26 (5) (2021) 2111–2122, https://doi.org/10.1007/s11036-021-01745-1.
[120] F. Tang, Y. Zhou, N. Kato, Deep reinforcement learning for dynamic uplink/downlink resource allocation in high mobility 5G HetNet, IEEE J. Sel. Area. Commun. 38 (12) (2020) 2773–2782, https://doi.org/10.1109/JSAC.2020.3005495.
[121] S. Ali, A. Haider, M. Rahman, M. Sohail, Y. Bin Zikria, Deep learning (DL) based joint resource allocation and RRH association in 5G-multi-tier networks, IEEE Access 9 (2021) 118357–118366 https://doi.org/10.1109/ACCESS.2021.3107430, Dl.
[122] A. Mohammed Mikaeil, W. Hu, L. Li, Joint allocation of radio and fronthaul resources in multi-wavelength-enabled C-RAN based on reinforcement learning, J. Lightwave Technol. 37 (23) (2019) 5780–5789, https://doi.org/10.1109/JLT.2019.2939169.
[123] D. Lynch, T. Saber, S. Kucera, H. Claussen, M. O'Neill, Evolutionary learning of link allocation algorithms for 5G heterogeneous wireless communications networks, GECCO 2019 - Proc. 2019 Genet. Evol. Comput. Conf. 1 (2019) 1258–1265, https://doi.org/10.1145/3321707.3321853.
[124] J. Mei, X. Wang, K. Zheng, An intelligent self-sustained RAN slicing framework for diverse service provisioning in 5G-beyond and 6G networks, Intell. Converg. Networks 1 (3) (2021) 281–294, https://doi.org/10.23919/icn.2020.0019.
[125] M. Vincenzi, E. Lopez-Aguilera, E. Garcia-Villegas, Timely admission control for network slicing in 5G with machine learning, IEEE Access 9 (2021) 127595–127610, https://doi.org/10.1109/ACCESS.2021.3111143.
[126] M. Alsenwi, S.R. Pandey, Y.K. Tun, K.T. Kim, C.S. Hong, A chance constrained based formulation for dynamic multiplexing of embb-urllc traffics in 5g new radio, Int. Conf. Inf. Netw. (2019) 108–113 https://doi.org/10.1109/ICOIN.2019.8718159, 2019-Janua.
[127] S.R. Pandey, M. Alsenwi, Y.K. Tun, C.S. Hong, A downlink resource scheduling strategy for URLLC traffic, in: 2019 IEEE Int. Conf. Big Data Smart Comput. BigComp 2019 - Proc., 2019, pp. 1–6, https://doi.org/10.1109/BIGCOMP.2019.8679266.
[128] Y. Li, C. Hu, J. Wang, M. Xu, Optimization of URLLC and eMBB multiplexing via deep reinforcement learning, in: 2019 IEEE/CIC Int. Conf. Commun. Work. China, ICCC Work. 2019, 2019, pp. 245–250, https://doi.org/10.1109/ICCChinaW.2019.8850168.
[129] A. Filali, Z. Mlika, S. Cherkaoui, A. Kobbane, Dynamic SDN-based radio access network slicing with deep reinforcement learning for URLLC and eMBB services, IEEE Trans. Netw. 9 (4) (2022) 2174–2187, https://doi.org/10.1109/TNSE.2022.3157274.
[130] Y. Huang, S. Li, C. Li, Y.T. Hou, W. Lou, A deep-reinforcement-learning-based approach to dynamic eMBB/URLLC multiplexing in 5G NR, IEEE Internet Things J. 7 (7) (2020) 6439–6456, https://doi.org/10.1109/JIOT.2020.2978692.
[131] J. Zhang, X. Xu, K. Zhang, B. Zhang, X. Tao, P. Zhang, Machine learning based flexible transmission time interval scheduling for eMBB and uRLLC coexistence scenario, IEEE Access 7 (2019) 65811–65820, https://doi.org/10.1109/ACCESS.2019.2917751.
[132] M. Almekhlafi, M. Chraiti, M.A. Arfaoui, C. Assi, A. Ghrayeb, A. Alloum, A downlink puncturing scheme for simultaneous transmission of URLLC and Embb traffic by exploiting data similarity, IEEE Trans. Veh. Technol. 70 (12) (2021) 13087–13100, https://doi.org/10.1109/TVT.2021.3116432.
[133] S. Khan, S. Khan, Y. Ali, M. Khalid, Z. Ullah, S. Mumtaz, Highly accurate and reliable wireless network slicing in 5th generation networks: a hybrid deep learning approach, J. Netw. Syst. Manag. 2 (2021), https://doi.org/10.1007/s10922-021-09636-2.
[134] M. Elsayed, M. Erol-Kantarci, Radio resource and beam management in 5G mmWave using clustering and deep reinforcement learning, in: 2020 IEEE Glob. Commun. Conf. GLOBECOM 2020 - Proc., 2020, https://doi.org/10.1109/GLOBECOM42002.2020.9322401.
[135] M.Y. Abdelsadek, Y. Gadallah, M.H. Ahmed, Resource allocation of URLLC and eMBB mixed traffic in 5G networks: a deep learning approach, in: 2020 IEEE Glob. Commun. Conf. GLOBECOM 2020 - Proc., 2020, https://doi.org/10.1109/GLOBECOM42002.2020.9322163.
[136] N. Salhab, R. Langar, R. Rahim, 5G network slices resource orchestration using Machine Learning techniques, Comput. Network. 188 (2021), 107829 https://doi.org/10.1016/j.comnet.2021.107829, August 2020.
[137] M. Alsenwi, E. Lagunas, S. Chatzinotas, Coexistence of eMBB and URLLC in Open Radio Access Networks : A Distributed Learning Framework, 2022, pp. 4601–4606.
[138] M. Zambianco, G. Verticale, A reinforcement learning agent for mixed-numerology interference-aware slice spectrum allocation with non-deterministic and deterministic traffic, Comput. Commun. 189 (March) (2022) 100–109, https://doi.org/10.1016/j.comcom.2022.03.010.
[139] M. Setayesh, S. Bahrami, V.W.S. Wong, Resource slicing for eMBB and URLLC services in radio access network using hierarchical deep learning, IEEE Trans. Wireless Commun. 21 (11) (2022) 8950–8966, https://doi.org/10.1109/TWC.2022.3171264.
[140] R. Kassab, O. Simeone, P. Popovski, Coexistence of URLLC and eMBB services in the C-RAN uplink: an information-theoretic study, in: 2018 IEEE Global Communications Conference, GLOBECOM 2018 - Proceedings, 2018, pp. 1–6, https://doi.org/10.1109/GLOCOM.2018.8647460.
[141] S.A. AlQahtani, W.A. Alhomiqani, A multi-stage analysis of network slicing architecture for 5G mobile networks, Telecommun. Syst. 73 (2) (2020) 205–221, https://doi.org/10.1007/s11235-019-00607-2.
[142] M. Ahsan, et al., Efficient Network Slicing for 5G Services in Cloud Fog-RAN Deployment over WDM Network, 2023, pp. 1–14, https://doi.org/10.1109/TVT.2023.3266234.
[143] B. Liu, P. Zhu, S. Member, J. Li, D. Wang, Energy-efficient optimization in distributed massive MIMO systems for slicing eMBB and URLLC services, IEEE Trans. Veh. Technol. (2023) 1–15 https://doi.org/10.1109/TVT.2023.3260988, PP.
[144] M. Setayesh, S. Bahrami, V.W.S. Wong, Joint PRB and power allocation for slicing eMBB and URLLC services in 5G C-RAN, 2020 IEEE Glob. Commun. Conf. GLOBECOM 2020 - Proc. (2020), https://doi.org/10.1109/GLOBECOM42002.2020.9322568.
[145] A. De Domenico, Y. Liu, W. Yu, Optimal computational resource allocation and network slicing deployment in 5G hybrid C-RAN, ICC 2019 - 2019 IEEE Int. Conf. Commun. (2019) 1–6.
[146] F. Kooshki, A.N.A.G. Armada, S. Member, Radio Resource Management Scheme for URLLC and eMBB Coexistence in a Cell-Less Radio Access Network, 11, 2023 March.
[147] M.N. Belaid, V. Audebert, B. Deneuville, R. Langar, Smart grid critical traffic routing and link scheduling in 5G IAB networks, in: 2022 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids, SmartGridComm, 2022, 2022, pp. 76–82, https://doi.org/10.1109/SmartGridComm52983.2022.9961009.
[148] M. Chen, Y. Miao, H. Gharavi, L. Hu, I. Humar, Intelligent traffic adaptive resource allocation for edge computing-based 5G networks, IEEE Trans. Cogn. Commun. Netw. 6 (2) (2020) 499–508, https://doi.org/10.1109/TCCN.2019.2953061.
[149] K. Boutiba, A. Ksentini, B. Brik, Y. Challal, A. Balla, NRflex : enforcing network slicing in 5G new radio, Comput. Commun. 181 (2022) 284–292 https://doi.org/10.1016/j.comcom.2021.09.034, September 2021.
[150] Y. Chen, C. Hsu, H. Hung, Optimizing communication and computational resource allocations in network slicing using twin-GAN-Based DRL for 5G hybrid C-RAN, Comput. Commun. 200 (2023) 66–85 https://doi.org/10.1016/j.comcom.2023.01.002, April 2022.
[151] R. Kassab, O. Simeone, P. Popovski, T. Islam, Non-orthogonal multiplexing of ultra-reliable and broadband services in fog-radio architectures, IEEE Access 7 (2019) 13035–13049 https://doi.org/10.1109/ACCESS.2019.2893128, Dl.
[152] R.K. Nandan, N.B. Adhikari, A multi-connectivity framework and simulation analysis of ultra-reliable low latency communication (URLLC) in 5G network, J. Inst. Eng. Ser. B 102 (5) (2021) 895–902, https://doi.org/10.1007/s40031-021-00600-x.
[153] Y. Fu, et al., Energy-efficient offloading and resource allocation for mobile edge computing enabled mission-critical internet-of-things systems, EURASIP J. Wirel. Commun. Netw. 2021 (2021) https://doi.org/10.1186/s13638-021-01905-7, 1.
[154] M. Setayesh, S. Bahrami, V.W.S. Wong, Joint PRB and power allocation for slicing eMBB and URLLC services in 5G C-RAN, 2020 IEEE Glob. Commun. Conf. GLOBECOM 2020 - Proc. (2020) 5–10, https://doi.org/10.1109/GLOBECOM42002.2020.9322568.
[155] A. Matera, Non-Orthogonal eMBB-URLLC Radio Access for Cloud Radio Access Networks with Analog Fronthauling, 2018, pp. 1–27, https://doi.org/10.3390/e20090661.
[156] J. Yun, Y. Goh, W. Yoo, 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. 9 (20) (2022) 20733–20749,

https://doi.org/10.1109/JIOT.2022.3177425.
[157] A. Dogra, R.K. Jha, S. Jain, A survey on beyond 5G network with the advent of 6G: architecture and emerging technologies, IEEE Access 9 (2021) 67512–67547, https://doi.org/10.1109/ACCESS.2020.3031234.
[158] B. Jamil, H. Ijaz, M. Shojafar, K. Munir, R. Buyya, "Resource allocation and task scheduling in fog computing and internet of everything environments : a taxonomy, Review , and Future Directions," 54 (2022) https://doi.org/10.1145/3513002, 11.
[159] A. Mamadou Mamadou, J. Toussaint, G. Chalhoub, Survey on wireless networks coexistence: resource sharing in the 5G era, Mobile Network. Appl. 25 (5) (2020) 1749–1764, https://doi.org/10.1007/s11036-020-01564-w.
[160] B.S. Khan, S. Jangsher, A. Ahmed, A. Al-Dweik, URLLC and eMBB in 5G industrial IoT: a survey, IEEE Open J. Commun. Soc. 3 (May) (2022) 1134–1163, https://doi.org/10.1109/OJCOMS.2022.3189013.
[161] I. Parvez, A. Rahmati, I. Guvenc, A.I. Sarwat, H. Dai, A survey on low latency towards 5G: RAN, core network and caching solutions, IEEE Commun. Surv. Tutorials 20 (4) (2018) 3098–3130, https://doi.org/10.1109/COMST.2018.2841349.
[162] R. Ali, Y. Bin Zikria, A.K. Bashir, S. Garg, H.S. Kim, URLLC for 5G and beyond: requirements, enabling incumbent technologies and network intelligence, IEEE Access 9 (2021) 67064–67095, https://doi.org/10.1109/ACCESS.2021.3073806.
[163] R. Yin, Z. Zou, C. Wu, J. Yuan, X. Chen, Distributed spectrum and power allocation for D2D-U networks: a scheme based on NN and federated learning, Mobile Network. Appl. 26 (5) (2021) 2000–2013, https://doi.org/10.1007/s11036-021-01736-2.
[164] S.K. Sharma, I. Woungang, A. Anpalagan, S. Chatzinotas, Toward tactile internet in beyond 5G era: recent advances, current issues, and future directions, IEEE Access 8 (i) (2020) 56948–56991, https://doi.org/10.1109/ACCESS.2020.2980369.
[165] S.R. Pokhrel, J. Ding, J. Park, O.S. Park, J. Choi, Towards enabling critical mMTC: a review of URLLC within mMTC, IEEE Access 8 (2020) 131796–131813, https://doi.org/10.1109/ACCESS.2020.3010271.
[166] F.S. Samidi, N.A.M. Radzi, W.S.H.M.W. Ahmad, F. Abdullah, M.Z. Jamaludin, A. Ismail, 5G new radio: dynamic time division duplex radio resource management approaches, IEEE Access 9 (2021) 113850–113865, https://doi.org/10.1109/ACCESS.2021.3104277.
[167] F. yan Tian, X. ming Chen, Multiple-antenna techniques in nonorthogonal multiple access: a review, Front. Inf. Technol. Electron. Eng. 20 (12) (2019) 1665–1697, https://doi.org/10.1631/FITEE.1900405.
[168] T. Akhtar, C. Tselios, I. Politis, Radio resource management: approaches and implementations from 4G to 5G and beyond, Springer US 27 (1) (2021), https://doi.org/10.1007/s11276-020-02479-w.
[169] M.O. Nait Belaid, V. Audebert, B. Deneuville, R. Langar, SD-RAN based approach for smart grid critical traffic routing and scheduling in 5G mobile networks, in: GLOBECOM 2022 - 2022 IEEE Global Communications Conference, 2022, pp. 5874–5879, https://doi.org/10.1109/GLOBECOM48099.2022.10000861.
[170] M. Bilal, V. Shanmuganathan, S. Won, Injecting cognitive intelligence into beyond-5G networks : a MAC, Comput. Electr. Eng. 108 (2023), 108717 https://doi.org/10.1016/j.compeleceng.2023.108717, December 2022.
[171] S. Hakak, et al., Autonomous vehicles in 5G and beyond : a survey, Veh. Commun. 39 (2023) 100551, https://doi.org/10.1016/j.vehcom.2022.100551.
[172] A. Kandoi, M. Raftopoulou, R. Litjens, Assessment of 5G RAN features for integrated services provisioning in smart cities, in: 2022 18th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), 2022, pp. 81–87, https://doi.org/10.1109/WiMob55322.2022.9941612.
[173] Y. Tang, S. Dananjayan, C. Hou, Q. Guo, S. Luo, Y. He, A survey on the 5G network and its impact on agriculture : challenges and opportunities, Comput. Electron. Agric. 180 (2021), 105895 https://doi.org/10.1016/j.compag.2020.105895, November 2020.
[174] W. Zhang, M. Derakhshani, S. Lambotharan, Stochastic optimization of URLLC-eMBB joint scheduling with queuing mechanism, IEEE Wirel. Commun. Lett. 10 (4) (2021) 844–848, https://doi.org/10.1109/LWC.2020.3046628.
[175] 5G Americas White Paper, *New Services & Applications with 5G Ultra Reliable low Latency Communications*, 2019, pp. 1–60, https://doi.org/10.31826/9781463236984-toc.
[176] K. Xiao, et al., Flexible multiplexing mechanism for coexistence of URLLC and eMBB services in 5G networks, 2020 ITU Kaleidosc. Ind. Digit. Transform. ITU K 19 (2) (2020) 82–90 https://doi.org/10.23919/ITUK50268.2020.9303213, 2020.
[177] H. Ji, S. Park, J. Yeo, Y. Kim, J. Lee, B. Shim, Ultra-reliable and low-latency communications in 5G downlink: physical layer aspects, IEEE Wireless Commun. 25 (3) (2018) 124–130, https://doi.org/10.1109/MWC.2018.1700294.
[178] J. Martyna, Heuristic design algorithm for scheduling of URLLC and eMBB traffics in 5G cellular networks, in: Internet of Things, Smart Spaces, and Next Generation Networks and Systems, 2022, pp. 438–448.
[179] X. Zhang, X. Guo, H. Zhang, RB allocation scheme for eMBB and URLLC coexistence in 5G and beyond, Wireless Commun. Mobile Comput. 2021 (2021), https://doi.org/10.1155/2021/6644323.
[180] R. Kobayashi, Y. Yuda, K. Higuchi, NOMA-based highly-efficient low-latency HARQ method for URLLC, in: 2021 IEEE 94th Vehicular Technology Conference (VTC2021-Fall), 2021, pp. 1–6, https://doi.org/10.1109/VTC2021-Fall52928.2021.9625306.
[181] V.D.P. Souto, S. Montejo-Sanchez, J.L. Rebelatto, R.D. Souza, B.F. Uchoa-Filho, IRS-aided physical layer network slicing for URLLC and eMBB, IEEE Access 9 (2021) 163086–163098, https://doi.org/10.1109/ACCESS.2021.3133139.
[182] Z. Qu, Z. Liu, X. Ding, H. Cao, G. Zhang, Co-existence analysis on satellite-terrestrial integrated IMT system, Mobile Network. Appl. 24 (6) (2019) 1926–1936, https://doi.org/10.1007/s11036-019-01337-0.
[183] M.A. Siddiqi, H. Yu, J. Joung, 5G ultra-reliable low-latency communication implementation challenges and operational issues with IoT devices, Electron 8 (9) (2019) 1–18, https://doi.org/10.3390/electronics8090981.
[184] J. He, G. Zhao, L. Wang, X. Sun, L. Yang, Secrecy analysis of short-packet transmissions in ultra-reliable and low-latency communications, EURASIP J. Wirel. Commun. Netw. 2021 (2021) https://doi.org/10.1186/s13638-020-01862-7, 1.
[185] A. Martin, et al., Network resource allocation system for QoE-aware delivery of media services in 5G networks, IEEE Trans. Broadcast. 64 (2) (2018) 561–574, https://doi.org/10.1109/TBC.2018.2828608.
[186] M.H. Alsulami, Challenges facing the implementation of 5G, J. Ambient Intell. Hum. Comput. (2022) 0123456789, https://doi.org/10.1007/s12652-021-03397-1.
[187] 3GPP, Etsi TS 123 502 - V15.2.0 - 5G; procedures for the 5G system 0 (2018).
[188] H. Zhang, N. Liu, X. Chu, K. Long, A.H. Aghvami, V.C.M. Leung, Network slicing based 5G and future mobile networks: mobility, resource management, and challenges, IEEE Commun. Mag. 55 (8) (2017) 138–145, https://doi.org/10.1109/MCOM.2017.1600940.
[189] B.M. ElHalawany, O. Hashad, K. Wu, A.S. Tag Eldien, Uplink resource allocation for multi-cluster internet-of-things deployment underlaying cellular networks, Mobile Network. Appl. 25 (1) (2020) 300–313, https://doi.org/10.1007/s11036-019-01288-6.
[190] H. Song, X. Fang, L. Yan, Handover scheme for 5G C/U plane split heterogeneous network in high-speed railway, IEEE Trans. Veh. Technol. 63 (9) (2014) 4633–4646, https://doi.org/10.1109/TVT.2014.2315231.
[191] N. Akkari, N. Dimitriou, Mobility management solutions for 5G networks: architecture and services, Comput. Network. 169 (2020) 107082, https://doi.org/10.1016/j.comnet.2019.107082.
[192] X. Xu, H. Zhang, X. Dai, Y. Hou, X. Tao, P. Zhang, SDN based next generation Mobile Network with Service Slicing and trials, China Commun 11 (2) (2014) 65–77, https://doi.org/10.1109/CC.2014.6821738.
[193] Z. Li, H. Shariatmadari, B. Singh, M.A. Uusitalo, 5G URLLC: design challenges and system concepts, Proc. Int. Symp. Wirel. Commun. Syst. (2018) https://doi.org/10.1109/ISWCS.2018.8491078, 2018-Augus, no. August.
[194] A. Kushchazli, A. Ageeva, I. Kochetkova, P. Kharin, A. Chursin, S. Shorgin, Model of radio admission control for URLLC and adaptive bit rate eMBB in 5G network, CEUR Workshop Proc 2946 (2021) 74–84.
[195] J. Yuan, Q. Xiao, R. Yin, W. Qi, C. Wu, X. Chen, Unlicensed assisted ultra-reliable and low-latency communications, Mobile Network. Appl. (2022), https://doi.org/10.1007/s11036-022-02003-8.
[196] A. Manzoor, S.M. Ahsan Kazmi, S.R. Pandey, C.S. Hong, Contract-based scheduling of URLLC packets in incumbent Embb traffic, IEEE Access 8 (2020) 167516–167526, https://doi.org/10.1109/ACCESS.2020.3023128.
[197] F. Firyaguna, A. Bonfante, J. Kibiłda, N. Marchetti, Performance Evaluation of Scheduling in 5G-mmWave Networks under Human Blockage, 1–8, 2020 [Online]. Available: http://arxiv.org/abs/2007.13112.
[198] A. Yazar, H. Arslan, Flexible multi-numerology systems for 5G new radio, J. Mob. Multimed. 14 (4) (2018) 367–394, https://doi.org/10.13052/jmm1550-4646.1442.
[199] M. Zambianco, G. Verticale, Mixed-numerology interference-aware spectrum allocation for eMBB and URLLC network slices, in: 2021 19th Mediterranean Communication and Computer Networking Conference (MedComNet), 2021, pp. 1–8, https://doi.org/10.1109/MedComNet52149.2021.9501277.
[200] M. Bennis, M. Debbah, H.V. Poor, Ultrareliable and low-latency wireless communication: tail, risk, and scale, Proc. IEEE 106 (10) (2018) 1834–1853, https://doi.org/10.1109/JPROC.2018.2867029.
[201] S. Martiradonna, A. Abrardo, M. Moretti, G. Piro, G. Boggia, "Architecting RAN Slicing for URLLC: Design Decisions and Open Issues," Proc. - 2019 IEEE/ACM 23rd Int. Symp. Distrib. Simul. Real Time Appl, DS-RT, 2019, pp. 1–4 https://doi.org/10.1109/DS-RT47707.2019.8958689, 2019.
[202] S. Bakri, B. Brik, A. Ksentini, On using reinforcement learning for network slice admission control in 5G: offline vs. online, Int. J. Commun. Syst. 34 (7) (2021), https://doi.org/10.1002/dac.4757.
[203] H. Basilier, A. Jan Lemark, T. Centonza, Åsberg, Applied network slicing scenarios in 5G, Ericsson Tech. (2021) 1–12.

Appendix-I (List of commonly used acronyms)

Abbreviation Elaboration

5G: 5th Generation mobile technology
4G: 4th Generation mobile technology
3GPP: 3rd generation partnership project
eMBB: Enhanced mobile broadband
URLLC: Ultra-reliable low latency communication
mMTC: Massive machine type communication
E2E: End to end
TTI: Transmission time interval
QoS: Quality of Service
NR: New Radio
C-RAN: Cloud radio access network
SCS: Sub Carrier Spacing

OFDMA: Orthogonal frequency division multiple access
NOMA: Non- Orthogonal multiple access
PRB: Physical Resource Block
RE: Resource Element
CQI: Channel Quality Indicator
CSI: Channel State Information
PUCCH: Physical uplink control channel
SDAP: Service data adaptation protocol
QFI: QoS flow identifier
DRB: Data Radio Barriers
RQA: Reflective QoS attributes
GBR: Guaranteed bit rate
Non-GBR: Non -guaranteed bit rate
UE: User Equipment
UPF: User Plane Function
QCI: QoS class Identifier
RR: Round Robin
PF: Proportional Fair
QGBRA: QoS Guaranteed Resource Block Allocation
HARQ: Hybrid automatic repeat request
DM: Dynamic Multiplexing
OS: Orthogonal Slicing
CBG: Code Block Group
MDS: Max distance separable code
MIMO: Multiple input Multiple Output
OMA: Orthogonal multiple Access
SCA-DC: Successive Convex approximation with difference convex programming
PL: Preference list
SBS: serving base station
PSUM: Penalty successive upper bound minimization
FDRA: fairness based distributed resource allocation
SCA: side-channel attack
SCA-RA: Side-channel attack resource allocation
TCM: Trellis-Coded Modulation
BLER: Block error rate
TDD: Time Division Duplex
TM: Transportation Model
MCC: Minimum Cell Cost
MODI: Modified Distribution
MEAR: Minimum Expected Achieved Rate
EDS: Equally Distributed Scheduler
MBS: Matching Based Scheduler
PS: Punctured Scheduler
MUPS: Multi-User Preemptive Scheduler
RS: Random Scheduler
NSBPS: Null space-based preemptive scheduler
gNB: Next generation Base Station
SCA-DC: Successive Convex approximation with difference convex programming
RSU: Roadside Unit
SQP: Sequential Quadratic Programming
CAC: Cell Admission Control
SINR: signal-to-interference-noise-ratio
SCA-RA: Side-Channel attack resource allocation
CARAM: Channel aware resource allocation for multiple numerologies
MCS: Modulation and Coding Scheme
DMRS: Demodulation reference signal miss detection
CMDP: Constrained Markov decision process
DNN: Deep neural network
MMD: Max-Matching Diversity
SCPF: Share Constrained Proportionally Fair
LSTM: Long Short memory
MINLP: mixed-integer non-linear programming
TDD: Time Division Duplex
DRL: Deep Reinforcement Learning
RRH: Radio Resource Head
EE-JWSBA: Energy-efficient joint workload scheduling and BBU allocation algorithm