IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 8 March 2023, accepted 29 March 2023, date of publication 5 April 2023, date of current version 10 April 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3264592

SURVEY

A Survey of Scheduling in 5G URLLC and Outlook for Emerging 6G Systems

MD. EMDADUL HAQUE[1], (Member, IEEE), FAISAL TARIQ[2], (Senior Member, IEEE), MUHAMMAD R. A. KHANDAKER[3], (Senior Member, IEEE), KAI-KIT WONG[4], (Fellow, IEEE), AND YANGYANG ZHANG[5]

[1]Department of Information and Communication Engineering, University of Rajshahi, Rajshahi 6205, Bangladesh
[2]James Watt School of Engineering, University of Glasgow, G12 8QQ Glasgow, U.K.
[3]School of Engineering and Physical Sciences, Heriot-Watt University, EH14 4AS Edinburgh, U.K.
[4]Department of Electronic and Electrical Engineering, University College London, WC1E 6BT London, U.K.
[5]Kuang-Chi Institute of Advanced Technology, Shenzhen 518057, China

Corresponding author: Faisal Tariq (Faisal.Tariq@glasgow.ac.uk)

This work was supported in part by the Engineering and Physical Sciences Research Council (EPSRC) under Grant EP/T015985/1.

ABSTRACT Future wireless communication is expected to be a paradigm shift from three basic service requirements of 5th Generation (5G) including enhanced Mobile Broadband (eMBB), Ultra Reliable and Low Latency communication (URLLC) and the massive Machine Type Communication (mMTC). Integration of the three heterogeneous services into a single system is a challenging task. The integration includes several design issues including scheduling network resources with various services. Specially, scheduling the URLLC packets with eMBB and mMTC packets need more attention as it is a promising service of 5G and beyond systems. It needs to meet stringent Quality of Service (QoS) requirements and is used in time-critical applications. Thus through understanding of packet scheduling issues in existing system and potential future challenges is necessary. This paper surveys the potential works that addresses the packet scheduling algorithms for 5G and beyond systems in recent years. It provides state of the art review covering three main perspectives such as decentralised, centralised and joint scheduling techniques. The conventional decentralised algorithms are discussed first followed by the centralised algorithms with specific focus on single and multi-connected network perspective. Joint scheduling algorithms are also discussed in details. In order to provide an in-depth understanding of the key scheduling approaches, the performances of some prominent scheduling algorithms are evaluated and analysed. This paper also provides an insight into the potential challenges and future research directions from the scheduling perspective.

INDEX TERMS 5G, 6G, packet scheduling, joint scheduling, URLLC, eMBB, mMTC.

I. INTRODUCTION

The 5G wireless services include three key enabling services considering the demand of all aspects of life. The enhanced Mobile Broadband (eMBB) service targets to enhance the data rate up to 10 Giga bits per second (Gbps); massive Machine Type Communication (mMTC) aims at supporting the connectivity of billions of Internet of Things (IoT) devices; and Ultra Reliable and Low Latency Communication (URLLC) is the service class to support machine-critical applications [1]. Among the service classes URLLC is the most innovative as the future networks are anticipated to include a vast set of applications that rely on mission-critical type communication including tactile internet, industrial automation and intelligent transportation system that requires the unprecedented level of high reliability and low latency [2], [3].

There are a number of characteristics which make URLLC service fundamentally different from the traditional networks. 3rd Generation Partnership Project (3GPP) outlines the general requirements for URLLC [4]. URLLC generally has very short packet size in order to meet the ultra-low end-to-end

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

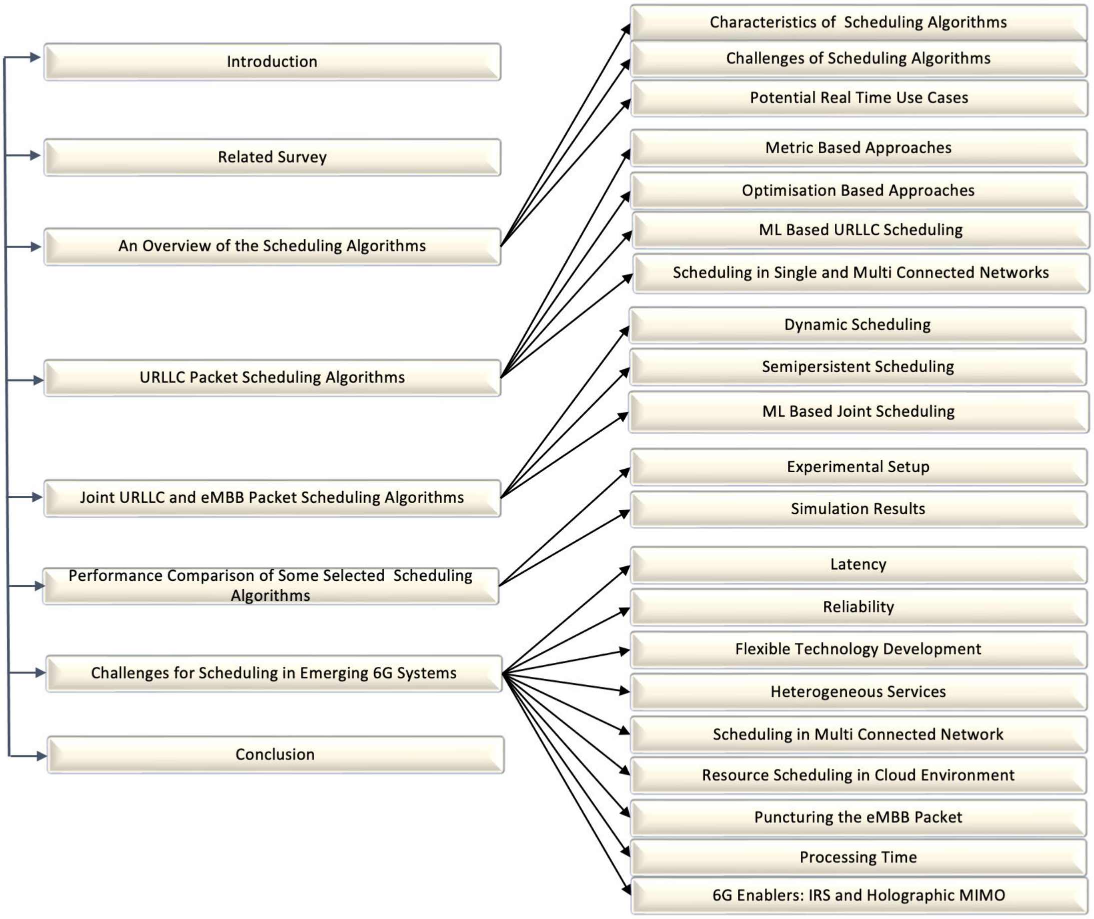

Introduction
Related Survey
An Overview of the Scheduling Algorithms
URLLC Packet Scheduling Algorithms
Joint URLLC and eMBB Packet Scheduling Algorithms
Performance Comparison of Some Selected Scheduling Algorithms
Challenges for Scheduling in Emerging 6G Systems
Conclusion
Characteristics of Scheduling Algorithms
Challenges of Scheduling Algorithms
Potential Real Time Use Cases
Metric Based Approaches
Optimisation Based Approaches
ML Based URLLC Scheduling
Scheduling in Single and Multi Connected Networks
Dynamic Scheduling
Semipersistent Scheduling
ML Based Joint Scheduling
Experimental Setup
Simulation Results
Latency
Reliability
Flexible Technology Development
Heterogeneous Services
Scheduling in Multi Connected Network
Resource Scheduling in Cloud Environment
Puncturing the eMBB Packet
Processing Time
6G Enablers: IRS and Holographic MIMO

FIGURE 1. **A tree diagram demonstrating organisation of the paper.**

(E2E) delay requirement. According to 3GPP specification the E2E delay for 5G is typically at most 1 ms. It is also aimed to ensure that the packets are received correctly with very high success probability in the range of $(1-10^{-5})$ to $(1-10^{-9})$. These stringent latency and reliability constraints are considered as the most challenging aspect of 5G network design. For 6 Generation (6G) E2E delay requirement is at most 0.1 ms. The peak data rate for eMBB of 5G and 6G are 20 Gbps and 1 Terra bits per second (Tbps), respectively. The effective radio resource utilisation is required for the system design considering the latency, reliability and data rate requirements.

The Radio Resource Management (RRM) utilised MAC and physical layer functionalities including resource sharing, Channel Quality Indicator (CQI) reporting, link adaptation through Adaptive Modulation and Coding (AMC), Hybrid Automatic Retransmission Request (HARQ) [5]. The effective radio resource allocation is crucial for the system. In reality, the efficient utilisation of radio resources is essential to meet the QoS requirements of URLLC.

The requirements of URLLC cannot be achieved with the existing Long Term Evolution (LTE) technology even with a single User Equipment (UE). The redesign of the basic components are required including Orthogonal Frequency Division Multiple Access (OFDMA) numerology, frame structure, Modulation and Coding Schemes (MCSs), HARQ, etc. [6]. Packet delivery time is a key consideration of the latency requirement of URLLC that mostly depends on scheduling policy. Packet scheduling is associated with 5G or 6G NodeB (gNB) which allocates Resource Block (RB) among the users by following specific strategies. Various approaches are used to schedule the packets in URLLC including Modified Largest Weighted Delay First (M-LWDF), Frame Level Scheduler (FLS), etc. [5].

In 5G and beyond system dedicated gNBs for URLLC UEs are not always possible. Thus coexistence of URLLC, eMBB and mMTC UEs are very likely under the same gNB. In this case, the same radio spectrum is shared among the services. Due to a hard latency boundary when all resources of a gNB is allocated for eMBB service and receives an URLLC packet, it suspends the ongoing eMBB transmission to free the radio resource for the URLLC packet. To maintain the QoS the immediate forwarding of the URLLC packets is proposed [7] which is called the puncturing mechanism [8].

In the technique eMBB transmission is immediately blocked and rescheduled after the URLLC transmission.

Considering the conventional scheduling policy and most recently explored aspects this paper categorises the packet scheduling algorithms into three categories including decentralised, centralised and joint scheduling. The decentralised scheduling section starts from the traditional LTE packet scheduling algorithm and covers the recent pioneer URLLC scheduling algorithms. The centralised scheduling algorithm section includes overview of the works from single connected and multi connected network perspectives. As the 5G and beyond systems multiple services are provided by the same gNB, the prominent joint scheduling algorithms are summarised in the next section. In addition to the literature survey an investigation of performance of the prominent algorithms is also presented.This paper also includes a detailed description about the research directions and future challenges of the scheduling algorithms.

There have been a number of works done on scheduling algorithms in wireless networks. Most of the surveys explore the scheduling algorithms on LTE networks. However, to the best of our knowledge, there is a lack of comprehensive investigation on scheduling algorithms in 5G and beyond systems. The main contributions of this paper are summarised as follows:

- An detailed discussion of existing packet scheduling algorithms for URLLC services.
- For a broader understanding this paper overviews the works under three categories including decentralised, centralised and joint scheduling algorithms.
- A summary of the algorithms are provided. The metrics of the metric based scheduling algorithms are also included in the Table 5.
- Potential research directions and future challenges in terms of scheduling mechanisms for URLLC in emerging 6G wireless systems.

The rest of the paper is organised as follows: Section II provides the description of related survey works found in literature. Section III provides a brief overview of the scheduling algorithms in terms of characteristics, challenges and potential real time use cases. Section IV provides the investigation of URLLC packet scheduling algorithms covering metric based approaches, optimisation based approaches as well as machine learning (ML) based approaches. This section also provides a comprehensive coverage of scheduling in single and multi connected networks. While section IV focuses on corer URLLC scheduling algorithms, Section V focuses on joint scheduling algorithms which includes both URLLC and eMBB traffic. The performance evaluation and analysis of some prominent metric based scheduling algorithms are covered in Section VI. Section VII offers a comprehensive discussion on the future research directions and potential challenges from scheduling perspective. Finally the paper is concluded in Section VIII. The Fig. 1 shows the structure of the paper and Table 1 provides a list of acronyms.

II. RELATED SURVEYS

There exists some survey paper on packet scheduling algorithms. The work in [9] provides a survey on content-aware downlink scheduling algorithms and radio resource allocation over LTE network. The paper provides a taxonomy that classifies the existing algorithms into two classes, namely context-aware and context-unaware. Further classification is also provided for detailed description of the algorithms. The existing scheduling techniques and the parameters are also listed to provide comparative understanding of the works. Finally, simulation results are presented to analyse the performance of some recent scheduling algorithms.

In [10] a survey of LTE scheduling algorithms and interference mitigation techniques is provided. Both downlink and uplink scheduling algorithms are investigated in the paper. The paper also surveys the interference mitigation techniques for the cell-edge users that face the high interference problem. Since the problem can be solved using enhanced frequency reuse techniques, the paper also investigates the enhanced techniques.

In [5] downlink scheduling algorithms are reviewed for LTE networks. The paper begins with a detailed overview of LTE networks and scheduling algorithms. It discusses the factors that should be considered before designing the protocols. For complete understanding the existing works the paper classifies them into five categories including (i) channel unaware, (ii) channel-aware/QoS-unaware, (iii) channel-aware/QoS-aware, (iv) semi-persistent for Voice over Internet Protocol (VoIP) support and (v) energy-aware. Pros and cons of the existing algorithms are discussed in the survey. The paper concludes with future research directions considering the evolution of communication technologies. The paper also highlights some important future challenges of the scheduling algorithms.

The authors in [11] provide a study on uplink scheduling algorithms for LTE from an Machine to Machine (M2M) perspective. The algorithms are classified based on different aspects including energy efficiency, QoS support, multi-hop transmission and scalability of the networks. A background on each category is presented in the paper. Then the detailed investigation of each category is provided.

In [12], a survey on uplink scheduling algorithms for LTE and LTE-Advanced are provided. The paper classifies the existing works into three categories including best effort schedulers, QoS based schedulers and power optimised schedulers. An evaluation framework is also presented in the paper.

While a number of survey works are found in literature on LTE, there are very few investigations that could be found on 5G or 6G. For example, [4] presents the detailed survey of URLLC. In the paper the authors present a brief overview of the 5G scheduling algorithms. Only URLLC scheduling algorithms are investigated excluding other two services including eMBB and mMTC. The authors in [13] present a survey of antenna and user scheduling techniques for massive Multiple Input Multiple Output (MIMO) 5G (MIMO-5G) wireless

TABLE 1. List of acronyms.

Acronym	Elaboration	Acronym	Elaboration
3GPP	3rd Generation Partnership Project	M-LWDF	Modified Largest Weighted Delay First
4G	4^{th} Generation	M2M	Machine to Machine
5G	5^{th} Generation	MAC	Medium Access Control
5GAA	5G Automotive Association	MCS	Modulation and Coding Scheme
5G-PPP	5G infrastructure Public Private Partnership	MDP	Markov Decision Process
6G	6^{th} Generation	MINLP	Mixed Integer Non-Linear Programming
AGV	Automated Guided Vehicles	MIMO	Multiple Input Multiple Output
AI	Artificial Intelligence	ML	Machine Learning
AMC	Adaptive Modulation and Coding	mMTC	massive Machine Type Communications
AoI	Age of Information	MR	Maximum Rate
AR	Augmented Reality	MTP	Motion-To-Photon
BCD	Block Coordinate Descent	MUPS	Multi-User Preemptive Scheduler
BET	Blind Equal Throughput	NAN	Neighbouring Area Network
BLS	Burst Limiting Shaper	NSBPS	Null-Space-Based Preemptive Scheduler
CoMP	Coordinated Multipoint	OFDMA	Orthogonal Frequency Division Multiple Access
CQI	Channel Quality Indicator	PF	Proportionally Fair
C-RAN	Centralised Radio Access Network	PGW	Packet data network Gateway
CSI	Channel State Information	PRB	Physical Resource Block
DL	Deep Learning	QoS	Quality of Service
DQN	Deep Q-Networks	RA	Resource Allocation
DRL	Deep Reinforcement Learning	RB	Resource Block
E2E	End to End	REG	Resource Regulator
EC	Effective Capacity	RF	Random Forest
eMBB	enhanced Mobile BroadBand	RF-ETDA	Random Forest based Ensemble TTI Decision Algorithm
eNB	evolved Node B	RRM	Radio Resource Management
EPC	Evolved Packet Core	SAFE-TS	Self-adaptive Flexible TTI Scheduling
EXP/PF	Exponential/Proportional Fair	SCA	Successive Convex Approximation
FIFO	First-In-First-Out	SG	Smart Grid
FLS	Frame Level Scheduler	SGNAN	Smart Grid Neighbour Area Network
Gbps	Giga bits per second	SLA	Service Level Agreement
gNB	6G NodeB	SRS	Split Responsibility Scheduler
GSA	Greedy Shrinking Algorithms	Tbps	Terra bits per second
HARQ	Hybrid Automatic Repeat reQuest	TTI	Transmission Time Interval
HMIMO	Holographic MIMO	THz	Terra Hertz
IRS	Intelligent Reflecting Surface	UE	User Equipment
IAB	Integrated Access and Backhaul	UMST	Ultra Mini Slot Transmission
INI	Intra-Numerology Interference	URLLC	Ultra Reliable Low Latency Communications
IoT	Internet of Things	USS	URLLC SLA Satisfaction
LIS	Large Intelligence Surface	V2X	Vehicle to Everything
LOG	Logarithmic	VoIP	Voice over Internet Protocol
LTE	Long Term Evolution	VR	Virtual Reality
LWDF	Largest Weighted Delay First	VUE	Vehicular User Equipment

TABLE 2. Summary of existing survey papers on scheduling techniques.

Paper Reference	URLLC	Generic Scheduling	URLLC Scheduling
M. M. Nasralla et al. (2018) [9]		✓	
R. Kwan et al. (2010) [10]		✓	
F. Capozzi et al. (2012) [5]		✓	
M. A. Mehaseb et al. (2015) [11]		✓	
N. Abu-Ali et al. (2013) [12]		✓	
G. J. Sutton et al. (2019) [4]	✓		
T. A. Sheikh et al. (2017) [13]		✓	
This Paper			✓

networks. Recent research works are investigated with key concentration of implementation detail of antenna and user scheduling algorithms in massive MIMO systems.

In [14] a survey of 6G wireless network presents a brief investigation of scheduling algorithms along with other aspects of 6G networks. The paper discusses the challenges and proposes some guidelines to mitigate the requirements of 6G networks. Another work on 6G network survey could be found in [15] with a brief investigation of scheduling algorithms. The paper presents some deployment scenarios of 6G networks and studies the performance of different scheduling strategies. The limitations of the application scenario are also presented in the paper. The authors in [16] present a survey on 6G wireless networks. The paper discusses the recent advancements and future trends in some aspects of 6G networks. A brief introductory survey of scheduling algorithms is also provided in the paper.

Although there have been some works focused on the packet scheduling algorithms, they are very early efforts and only provide surveys on LTE networks and many recent works are not covered. On the other hand, few works are available on the very introductory study on scheduling algorithms of 5G and beyond systems, but detailed overview on URLLC or joint scheduling algorithms with the three services including URLLC, eMBB and mMTC are not available. The existing surveys are summarised in Table 2.

To the best of our knowledge, comprehensive study on recent packet scheduling algorithms is still missing for 5G and beyond systems. To this end, this paper fills the gap by providing a state-of-the-art survey of scheduling algorithms, especially the joint scheduling algorithms along with the URLLC scheduling algorithms.

III. AN OVERVIEW OF THE SCHEDULING ALGORITHMS

The process of assigning the shared resources among the users at a given time with some performance guarantee is so-called packet scheduling. The task of scheduling algorithms is to determine which packet should be served first among the packets that are generally organised into a queue.

The Fig. 2 depicts a generic representation of a scheduler. When UE traffic arrives, the traffic classifier divides them into different queues according to their traffic type. Then the traffic prioritiser will act depending on the scheduling algorithm applied in the system. For example, if a scheduler wants URLLC traffic to have highest priority, the prioritiser will ensure that the URLLC traffic gets scheduled at first. The time is divided into slots in LTE networks. The resource allocation occurs in every time slot with 0.5 ms duration of each [5]. But in 5G the latency requirements are different for different services and applications as shown in Table 3 [17], [18], [19], [20]. For many applications, the latency requirement is 1 ms, hence with the LTE the E2E or round-trip latency of 1 ms is not practicable. In order to achieve the required latency, the slot is divided into mini-slots in 5G and beyond. A scheduler needs certain information during the scheduling decision including number of sessions, link state, the status of queue and head of line delay. For downlink scheduling, this information is required by base station. On the other hand, for uplink scheduling this information is required by the mobile station. While the base station can easily find the information, the mobile station needs some extra steps to get the information.

For delay-sensitive applications each packet in the queue contains the time information. That means the packs are time stamped and from the time information the scheduler computes the remaining lifetime of the packet. If the head of line delay of a packet i.e. the lifetime of the packet expires, then the packet is dropped.

If a transmission fails, then retransmission is a normal way to achieve the reliability [21] The retransmission can also fail. The policy and retransmission starting time affect the success in the retransmission. For example, in URLLC the E2E delay is less than the channel coherence time, thus successive retransmission may not be beneficial to achieve the reliability [21]. Different techniques have been found in the literature including retransmission with frequency hopping and pre-scheduling resources for retransmission [21], [22]. Thus, retransmission scheduling is an important part of the scheduling algorithm that affects the reliability as well as the latency of the packets.

A. CHARACTERISTICS OF SCHEDULING ALGORITHMS

A scheduling algorithm should have the following characteristics [5], [23]:

1) THROUGHPUT

Maximising the throughput is a primary goal of a scheduling algorithm. The algorithm should provide the guaranteed instantaneous or short-term maximum throughput during deep fading condition as well as long-term throughput.

2) EFFICIENT LINK UTILISATION

To achieve the effective utilisation the algorithm should be designed in a way that maximises the number of users served in a given time. That means a scheduler should not assign a transmission slot to a user with poor channel condition that has the high probability of transmission failure. A performance indicator, known as Goodput, is widely used for efficient link utilisation [5]. It is a measure of actual data rate excluding the overheads and packet retransmission due to physical layer error which gives better indication of useful data transfer rate..

3) DELAY BOUND

In order to satisfy the delay sensitive application the packet should schedule within a predefined time. Especially the scheduling algorithm must satisfy the stringent E2E delay bound requirement for URLLC packets.

4) FAIRNESS

Maximising the throughput ensures the efficient link utilisation. But it may cause unfair channel allocation among the UEs. Hence, fairness is one of the major requirements of a wireless system that must be taken into consideration during the design of a packet scheduling algorithm.

5) COMPLEXITY

A schedule makes decisions about resource allocation quite frequently. Usually it re-allocates resources among the UEs in every 1 ms interval. Thus a low complexity scheduling algorithm is required for the emerging high speed networks, specially for the URLLC service.

B. CHALLENGES OF SCHEDULING ALGORITHMS

The traditional communication systems such as LTE were designed to provide high data transmission rate and reliability. The LTE system can achieve high reliability in physical layer data transmission at the expense of high latency from

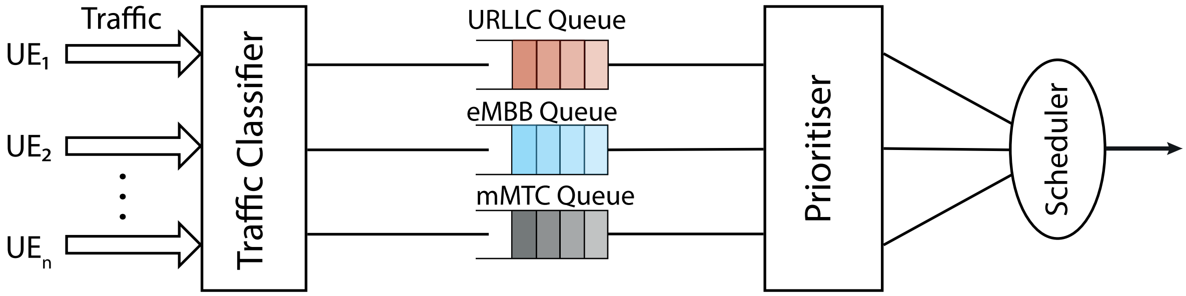


FIGURE 2. A typical scheduling mechanism of 5G wireless systems.

TABLE 3. Latency requirement of URLLC applications.

Services	Applications	Latency requirement	Reliability requirement
URLLC	Autonomous driving	1 ms	99.9999%
	Electricity distribution (medium voltage)	20 ms	99.9%
	Electricity distribution (high voltage)	5 ms	99.9999%
	IoT (Internet of things/Tactile Internet)	1 ms	99.99%
	Augmented/Virtual Reality	5 - 10 ms	99.9 % - 99%
eMBB	Public gatherings	10 - 100 ms	99%
	Shopping centers	10 - 100 ms	99%
	Mass communication	200 - 5000 ms	99%
mMTC	Industry 4.0, vehicles, Haptics	< 50 ms	99.9% - 99.99999%
	IIoT automation/orchestration	10 - 50 ms	99.9% - 99.99999%
	Vehicles, AR/VR, Drones	2 - 10 ms	99.9% - 99.999%
	Autonomous cars, Haptics	2 ms	99.99999%

tens to hundreds of milliseconds. This high latency is due to 1 ms Transmission Time Interval (TTI), large processing delay at both transmitter and receiver side, and the retransmission policy. Thus 1 ms latency is hardly possible in the current LTE networks where the time axis is divided into TTI of 1 ms. On the other hand, in LTE $1 - 10^{-1}$ (90% percent) reliability is the default setting which can be extended up to $1 - 10^{-5}$ with 3 HARQ retransmissions. But for each HARQ retransmission induces the delay of 8 TTIs. Hence the LTE systems cannot fulfil the requirements of 5G. To fulfil the requirement 3GPP proposed 5G networks with 14 times shorter TTI than that of LTE [24].

The main challenge of URLLC service is to achieve the 1 ms latency while providing the high reliability. These requirements depend on applications. The general version of requirements are as follows.

1) HIGH RELIABILITY

The general version of requirements are specified in [25]. One of the most important requirements is high reliability which is specified in the release as follows. The reliability requirement is specified as 99.999% for short message transmission.

2) LOW LATENCY

As discussed above, with the emergence of new applications requiring real time interventions and interactions, latency requirements are becoming increasingly stringent. In 3GPP Release 15, the user-plane and link level (both downlink and uplink) latency requirements are specified as 1 ms and 0.5 ms respectively.

To fulfil the requirement to the ever-growing number of users of 5G the efficient utilisation of the network resources is crucial for 5G design. Especially in the heterogeneous service environment where a huge amount of resources is required for eMBB traffic. Moreover, the huge amount of mMTC devices is also a factor of the network especially for stringent latency and reliability bounded URLLC service. Thus, scheduling the URLLC packets is a challenging task due to fundamental trade-off between the latency and reliability [26]. For instance, to satisfy the stringent requirements of URLLC the system should immediately transmit the URLLC packet by cancelling the ongoing other transmissions.

Categorising scheduling algorithms is a difficult task as each algorithm may possess multiple characteristics. Therefore, scheduling algorithms are broadly categorised here into two different categories. First, the algorithms that are primarily focused on latency and reliability issues of the packet delivery. These algorithms include 5G URLLC scheduling algorithms and are thoroughly described in section IV. The second category constitutes the algorithms which handles heterogeneous traffic types which schedules both URLLC, eMBB and mMTC traffics together and are referred to as joint scheduling algorithms which are described in the section V.

In recent years, various ML based techniques are increasingly adopted to improve efficiency of the algorithms. Due to it's growing prominence, ML algorithms for both URLLC and joint scheduling categories are presented separately in sections IV-C and V-C respectively.

The algorithms can also be categorised according to architecture of the scheduler such as centralised and decentralised. Since the majority of the algorithms are decentralised, they are not presented in a separate category. Instead, only centralised algorithms are specified and all decentralised algorithms are discussed in various categories such as metric based approaches and optimisation based approaches.

C. POTENTIAL REAL TIME USE CASES

A broad range of use cases could be found that need URLLC service. Some very important vertical use cases include Augmented Reality (AR) and Virtual Reality (VR), tactile internet, smart grid and factory automation.

1) AUGMENTED REALITY AND VIRTUAL REALITY

In many heavy industries the AR is required for a number of purposes including fault finding and repairing the fault. For example, in a nuclear power plant workers cannot intervene in the fault places in many cases. In this case the workers have to repair the fault from a distance location with the help of augmented reality. The worker can repair the fault remotely with the help of special software and different equipment. In this case the URLLC of a 5G network is a possible means of information between the worker and the actuator and the sensor. In this case the required latency and reliability are 10 ms and 99.9999%, respectively [27]. There should be many other requirements including real time data processing, high security in data transmission and high-rate video streaming. 360 panoramic VR is expected to be the next generation video broadcasting technology to experience the real time environment for being like in a stadium watching a game in any direction. In this case the 360 cameras are required to install in the location to provide the view from different angles at the same time. One of the key requirements of the application is the motion-to-photon (MTP) requirement of humans, which is 20ms [28]. Otherwise, the audience may face cybersickness. To eliminate the sickness the round-trip delay between the audience and video cameras should be less than 20 ms. The other requirements include 99.999% reliability, user experience throughput in between 40 Mbps to 5Gbps.

2) TACTILE INTERNET

A potential use case of tactile internet is in the robotic telesurgery system [29]. The tele-surgery provides the healthcare surgical service to the patients in remote locations. The system basically has three components including doctor's end, patient's end and the network part. The doctor operates a patient by controlling the robotic arms at the patient's end. The doctor's and patient's end are connected through the internet. Low latency data transfer is the primary concern for the system. There are some other issues including high reliability and data rate that are also important requirements. The requirements differ from application to application. For heart disease tele-surgery the latency must be less than 1ms and the reliability at least 99.999% [30].

3) SMART GRID

Smart grid is another important use case of 5G which is expected to become capable of generating and distributing electricity in an efficient, sustainable, reliable, flexible and secure manner. The smart grid is a sophisticated integration of energy grid and 5G. The grid provides real time monitoring and remote energy monitoring through smart metres. It has the capacity of fault detection, finding and correction. The grid also facilitates the renewal of energy generation and distribution. For the above functionality the smart grid needs the URLLC and mMTC services. For different parts of the grid the communication requirements are different. For example, the required data rates for grid access, grid backhaul and grid backbone are 1Kbps, several Mbps and up to several Gbps, respectively [31]. Latency requirements are <1s, <50ms and <5ms for grid access, grid backhaul and grid backbone, respectively. Maximum allowable packet losses are 10^{-6} and 10^{-9} for grid backhaul and grid backbone, respectively.

4) FACTORY AUTOMATION

Future factories are expected to be the integration of factory equipment and 5G networks. For heavy industry there will be the Automated Guided Vehicles (AGV) that will transport products, tools, raw materials from one location to another in the factory environment. Basically, the AGV are mobile robots that have the capacity of handling materials, monitoring and control, image processing, recognition etc. to do the complex tasks. The unmanned AGVs need different information from different sensors, actuators and AGVs with very high reliability and low latency to perform the tasks. In the factory automation use case the required latency, reliability and remote control bit rates are 5ms, 99.999% and 100kbps, respectively [27].

IV. URLLC PACKET SCHEDULING ALGORITHMS

In this section we have reviewed the benchmark packet scheduling algorithms starting from the preliminary algorithms of LTE to the most prominent algorithms that can be used for 5G wireless networks. The algorithms are categorised into metric based and optimised based scheduling algorithms. Various notations that are used to describe the algorithms are given in Table 4.

A. METRIC BASED APPROACHES

1) MAXIMUM RATE SCHEDULER

The aim of the Maximum Rate (MR) scheduling algorithm [32] is to maximise the system throughput without considering individual user performance. This algorithm does not guarantee a fair allocation among the users. In each assignment interval, the algorithm chooses a user that maximises

TABLE 4. Summary of notations used in scheduling matrix.

Expression	Meaning
$S^X_{i,t}$	Generic metric for X algorithm of i^{th} user in time interval t
$r_i(t)$	Data rate achieved by the i^{th} user at time t
$R_i(t)$	Past average throughput achieved by the i^{th} user up to time t
τ_i	Delay threshold for the i^{th} user
δ_i	Acceptable packet loss rate for the i^{th} user
D^{HL}_i	Head of line delay of the i^{th} user

the metric given below:

$$S^{MR}_{i,t} = \arg\max_i r_i(t). \tag{1}$$

Therefore, a user with relatively poor channel condition has less chance for transmission. If a user continuously experiences poor channel condition, then according to this algorithm it may not get scheduled at all which may also affect the overall system throughput. This is a channel aware but quality of service unaware scheduling algorithm. The algorithm performs best when all the users experience relatively good channel conditions.

2) BLIND EQUAL THROUGHPUT

In Blind Equal Throughput (BET) the resource is allocated considering the past average throughput [33]. At a particular time a UE with the lowest past average throughput is allocated the resource. The metric for BET is given below:

$$S^{BET}_{i,t} = \frac{1}{R_i(t)}, \tag{2}$$

where

$$R_i(t) = \beta R_i(t-1) + (1-\beta)R_i(t) \quad \text{and } 0 \leq \beta \leq 1. \tag{3}$$

From the metric, it is clear that the UE with lowest experience throughput in the past will be allocated resources as long as it does not reach higher than other UEs in the cell. If a transmitter frequently fails to transmit packets to UEs due to poor channel condition, It usually has a low past average throughput. Thus the UE with the poor channel condition will be served more often than others to ensure fairness. It doesn't take channel conditions in deciding the allocation and is categorised as a channel unaware scheduling algorithm. The past average throughput of i^{th} user at time t is represented by $R_i(t)$. This is defined as moving average throughput and is updated every TTI for each user. The metric indicates that BET has a significant policy to ensure the fairness in packet transmission. But if a user continuously experiences poor channel condition then the scheduler frequently allocates resources to the user as the transmission frequently fails due to the channel condition. This eventually degrades the network throughput.

3) PROPORTIONAL FAIR SCHEDULER

Proportional Fair (PF) scheduling algorithm [34] aims to strike a balance between maximising the system throughput and ensuring equal throughput for all the users in the system. PF scheduler is a channel aware but QoS unaware scheduler that combines the MR and BET algorithm [5] in the metric to make a balance between the two. The metric of PF is is given below:

$$S^{PF}_{i,t} = S^{MR}_{i,t} S^{BET}_{i,t} = \arg\max_i \left\{ r_i(t) \frac{1}{R_i(t)} \right\}, \tag{4}$$

where $R_i(t)$ represents the average throughput of the user i at the time interval t. $R_i(t)$ is evaluated as:

$$R_i(t) = \left(1 - \frac{1}{t_c}\right) R_i(t-1) + \frac{1}{t_c} r_i(t) \tag{5}$$

where t_c is a time constant. Eq. 4 indicates that the PF scheduler takes into consideration the number of data packets an individual user transmits over a defined time window. Higher the value of transmitted data rate of a user in the past window indicates lower the chance of the user to get scheduled. Therefore it ensures fairness among the users over a certain time duration. Several extended versions of the PF algorithm could be found in literature that consider different perspectives to increase the performance including [35], [36], [37], [38], [39], and [40]. PF scheduler shows better performance than MR and BET. However, PF is still not suitable in its current form for 5G or beyond systems since it does not consider the packet validity period in the latency bound service which is particularly required for URLLC.

4) EXPONENTIAL/PROPORTIONAL FAIR SCHEDULER

In [34], the authors propose an exponential rule based channel and QoS aware scheduler named Exponential/Proportional fair (EXP/PF) scheduling algorithm which was originally developed to support the multimedia applications [41]. The generic metric used in EXP/PF algorithm to compute the real time flow is as follows:

$$S^{EXP/PF}_{i,t} = \arg\max_i EXP \left\{ \frac{\alpha_i D^{HL}_i - W}{1 + \sqrt{W}} \right\} r_i(t) \frac{1}{R_i(t)}, \tag{6}$$

where,

$$W = \frac{1}{N_{af}} \sum_{i}^{N_{af}} \alpha_i D^{HL}_i, \tag{7}$$

and N_{af} is the number of active downlink real-time flows.

Logarithmic (LOG) and Exponential (EXP) rules are presented in [42] that considers the recent allocation history with low computational complexity. The metric for LOG rule is as follows:

$$S^{LOGrule}_{i,t} = \arg\max_i b_i log(c + a_i D^{HL}_i) K_i, \tag{8}$$

where a_i, b_i and c are fixed positive parameters; and K_i is channel spectral efficiency of i^{th} user.

EXP rule is similar to LOG rule to enhance the performance of PF. The metric for EXP rule is given below:

$$S_{i,t}^{EXPrule} = \arg\max_i EXP \left\{ \frac{\alpha_i D_i^{HL}}{c + \sqrt{(1/N)\sum_j D_i^{HL}}} \right\} K_i, \quad (9)$$

where N is the number of active users participating in downlink real-time flows.

The LOG rule metric increases the value logarithmically of the head of line delay on the other hand, EXP rule metric increases the metric exponentially over the head of line delay. However, EXP/PF, LOG rule and EXP rule scheduling algorithms do not consider the transmission latency bounding requirement of 5G network.

5) LARGEST WEIGHTED DELAY FIRST

Largest Weighted Delay First (LWDF) is a channel unaware scheduling algorithm designed for wired networks and applicable for real-time applications. It avoids deadline expiration by considering the head of line delay in the metric [43]. The LWDF scheduling algorithm is based on acceptable packet loss rate for the users. The loss rate is based on the deadline expiration. The metric of the scheduling algorithm is given below:

$$S_{i,t}^{LWDF} = \alpha_i D_i^{HL} \quad (10)$$

where

$$\alpha_i = -\frac{log\delta_i}{\tau_i} \quad (11)$$

In this case the user with highest priority in terms of the packet loss rate and deadline expiration is selected for the resource allocation. In the metric if HL of two flows are equal then α_i plays an important role.

6) MODIFIED LARGEST WEIGHTED DELAY FIRST SCHEDULER

The M-LWDF [44] is the modification of LWDF that combines LWDF and PF to improve the performance. In this algorithm the packet delivery delay is bounded. The metric of the algorithm is computed by multiplying the LWDF and PF metrics [45]. The metric of M-LWDF can be written as:

$$S_{i,t}^{M-LWDF} = \arg\max_i \alpha_i D_i^{HL} r_i(t). \frac{1}{R_i(t)} \quad (12)$$

where D_i^{HL} represents the head of line packet delay of user i. Now, the packet drop probability, α_i of the user is calculated as

$$\alpha_i = -\frac{\log \delta_i}{\tau_i} \quad (13)$$

where δ_i is acceptable packet loss rate for the i^{th} user and τ_i is the delay threshold for the i^{th} user. M-LWDF uses the accumulated delay for shaping the behaviour of PF, fairness and QoS provisioning [5].

7) COMBINED SCHEDULING ALGORITHM

An algorithm that combines the benefits of M-LWDF and the LOG rule is proposed in [46]. The algorithm provides better performance compared to its constituent individual algorithms. It is also suitable for URLLC applications. The algorithm is both channel and QoS aware. It uses weighted delay first as a logarithmic parameter and combines with M-LWDF metric. The logarithm of the parameter smooths the subcarrier selection criteria that improves the performance. The metric of the algorithm is defined as follows:

$$S_{i,t}^{(Prop)} = \arg\max_i \left\{ \log\left(1 + \alpha_i D_i^{HL}\right) S_{i,t}^{(M-LWDF)} \right\} \quad (14)$$

where, $S_{i,t}^{(Prop)}$ represents the generic metric of the i^{th} user in t time interval of the proposed algorithm, $S_{i,t}^{(M-LWDF)}$ is the generic metric of the i^{th} user in time interval t of M-LWDF algorithm which is calculated by the eq. 12, α_i is the packet drop probability of the i^{th} user due to the life time expiration which is calculated by the eq 16. D_i^{HL} is the head of line delay of the user i which is computed by subtracting the packet arrival time from the current time according following relation:

$$D_i^{HL} = t - t_a(i) \quad (15)$$

where $t_a(i)$ is the packet arrival time and α_i is defined as follows

$$\alpha_i = -\frac{\log \delta_i}{\tau_i}. \quad (16)$$

Note that for a particular packet of user i, if $D_i^{HL} \geq 1ms$ then the packet will be dropped according to the requirement of URLLC. The authors show that the performance of the combined scheduling algorithm is better than several benchmark scheduling algorithms but the computation cost is little higher. However, the summary of the metric based scheduling techniques are presented in Table 5.

B. OPTIMISATION BASED APPROACHES

Metric based method uses fixed formula based calculation for making a decision. On the other hand, optimisation based approaches maximise or minimise a real function. These algorithms take input values from an allowed set and then systematically use a method to compute the maximum or minimum value of the function.

Several research works are found in the literature on optimisation based approaches. For example, in [47] an optimisation based algorithm is proposed for downlink scheduling for URLLC. The paper converts the problem into the URLLC Service Level Agreement (SLA) Satisfaction (USS) problem. A dynamic programming technique is used to solve the problem for small instances. For large instances, the problem is NP-hard. To solve the large instance a low complexity near optimal knapsack-inspired heuristic is proposed. Although the paper finds near-optimal results, the QoS issues in terms of packet loss rate and variable block length URLLC transmissions are not considered [48], [49].

TABLE 5. Summary of the metric based scheduling algorithms.

Scheduling algorithms	Metric	Generation	Candidate of URLLC?
MR [32]	$S_{i,t}^{MR} = \arg\max_i r_i(t)$	4G	No
BET [33]	$S_{i,t}^{BET} = \frac{1}{R_i(t)}$, where $R_i(t) = \beta R_i(t-1) + (1-\beta)R_i(t)$ and $0 \leq \beta \leq 1$	4G	No
PF [34]	$S_{i,t}^{PF} = \arg\max_i \left\{ r_i(t) \frac{1}{R_i(t)} \right\}$, where $R_i(t) = \left(1 - \frac{1}{t_c}\right) R_i(t-1) + \frac{1}{t_c} r_i(t)$	4G	Yes
EXP/PF [34]	$S_{i,t}^{EXP/PF} = \arg\max_i EXP \left\{ \frac{\alpha_i D_i^{HL} - W}{1+\sqrt{W}} \right\} r_i(t) \frac{1}{R_i(t)}$, where $W = \frac{1}{N_{af}} \sum_i^{N_{af}} \alpha_i D_i^{HL}$	4G	Yes
LOG rule [42]	$S_{i,t}^{LOGrule} = \arg\max_i b_i log(c + a_i D_i^{HL}) K_i$,	4G	Yes
EXP rule [42]	$S_{i,t}^{EXPrule} = \arg\max_i EXP \left\{ \frac{\alpha_i D_i^{HL}}{c+\sqrt{(1/N) \sum_j D_i^{HL}}} \right\} K_i$,	LTE	Yes
LWDF [43]	$S_{i,t}^{LWDF} = \alpha_i D_i^{HL}$, where $\alpha_i = -\frac{log\delta_i}{\tau_i}$	4G	Yes
M-LWDF [44]	$S_{i,t}^{M-LWDF} = \arg\max_i \alpha_i . D_i^{HL} . r_i(t) . \frac{1}{R_i(t)}$	5G	Yes
Modified efficient M-LWDF [46]	$S_{i,t}^{(Prop)} = \arg\max_i \left\{ \log\left(1 + \alpha_i \cdot D_i^{HL}\right) \cdot S_{i,t}^{(MLWDF)} \right\}$, where $D_i^{HL} = t - t_a(i)$	5G	Yes

To take the URLLC packet scheduling decision, 5G networks use low complexity computation. Thus the base station may not acquire accurate Channel State Information (CSI) due to low latency requirements of URLLC. To address the issue, [50] explores frequency diversity where a packet will be simultaneously sent over multiple channels. The base station gets the information of the channels from the feedback of the received signal by the UE. The base station uses the feedback to address the problem of dynamic channel allocation for URLLC service in absence of accurate CSI. The Markov decision process framework is used to formulate the problem. Then a low complexity approximation algorithm is proposed for the problem. However, the paper does not study the performance with multi-antenna and multiuser URLLC transmission [51].

A joint admission control [52], [53] and resource scheduling algorithm is proposed in [54] for URLLC in 5G network. The paper considers the continuous and binary models for resource scheduling. Scheduling the URLLC traffic for a particular number of users in both models is NP-complete. Finally, the paper finds an approximation algorithm for any number of URLLC users. However, the paper shows the performance of the algorithm without considering the realistic scenarios including multi-cell setting and heterogeneous real-time requirements of links [55].

Multi-cell and multi-channel URLLC scheduling technique for 5G and beyond networks is proposed in [55]. Instead of prior studies on probabilistic per-packet reliability and latency guarantee for single cell and single channel network [56], the paper [57] considers the reliability and latency guarantee for multi-cell and multi-channel network which is more realistic. The paper proposes a distributed local-deadline-partition scheduling algorithm for the realistic environment. The algorithm shows effective performance in terms of the number of URLLC traffic served. That means the method can support more URLLC traffic in a real-time environment than traditional algorithms.

A differential QoS oriented scheduling algorithm under given delay tolerance is proposed in [58]. First-In-First-Out (FIFO) service discipline based queue is used to model the system which is represented by a Markov chain. This is the extension of the authors previous work [59], [60] for multiple services. However, to reduce the searching space the states of similar behaviour in the chain are aggregated. This policy reduces significantly the complexity of the algorithm. However, the FIFO discipline is suitable for the application of homogeneous latency requirement based application. But for heterogeneous latency requirement applications the discipline is not applicable [61].

A robust packet scheduling for OFDM system is proposed in [62]. An optimisation technique is used to minimise the Physical Resource Block (PRB) assignment and power allocation under the required delay and reliability constraints. A low complexity successive convex approximation based method is used to solve the problem that yields the sub-optimal solution. Although the technique shows efficient performance in a simplified environment, a realistic scenario with a multi-cell environment is required to evaluate the performance study.

An uplink resource scheduling technique for Smart Grid (SG) Neighbourhood Area Network (SGNAN) is proposed in [63]. SGNAN consists of intelligent household electrical appliances and the home gateway that collects data and sends it to the data concentrator unit of the NAN. Different delay, bandwidth and reliability are required for different home area networks including residential, business and industry. To meet the different quality demands the paper proposes the priority based URLLC scheduling algorithm for the SG.

A similar work of resource scheduling in heterogeneous cellular networks for SG is found in [64]. The optimisation technique is used to maximise the system throughput and first-order Taylor expansion is used to approximate the solution. However, different latencies are required for different networks of the SG. Hence the results with variable latency are required in the study. The key features of the optimisation based techniques are summarised in Table 6.

C. ML BASED URLLC SCHEDULING

A risk-aware machine learning based algorithm is proposed in [65] for the coexistence of scheduled and non-scheduled URLLC traffic. The paper proposes a hybrid orthogonal/non-orthogonal multiple access scheme for the coexistence scenario. Distributed risk-aware ML based solution proposed radio resource management to increase the efficiency of the network.

A machine learning based resource scheduling problem is proposed in [66] to achieve the timeliness in remote factory monitoring. In particular, a reinforcement learning algorithm is proposed in the paper to minimise age of information in URLLC. An optimization problem is formulated with an integer non-convex constraint which is most unlikely to solve in polynomial time. The authors used a reinforcement learning technique to solve the problem.

A scheduling and resource allocation technique is proposed in [67] to increase the reliability and extra protection in URLLC users and cell edge users against Intra-Numerology Interference (INI). The INI is the additional interference of different numerology as adopted by 3GPP. The INI aware scheduling provides more resources to URLLC to ensure the QoS. It also provides more resources to the cell edge URLLC users.

D. SCHEDULING IN SINGLE AND MULTI CONNECTED NETWORKS

Single-connectivity refers to the network configuration in which a UE is connected to the network usually through a single base station using a single connection. On the other hand, in a multi connected network, the UE is simultaneously connected to the network through multiple base stations. In the multi-connected network the same data is transmitted to multiple base stations. This increases the reliability of the transmission which is required for 5G network in general and for URLLC in particular.

1) SCHEDULING IN SINGLE CONNECTED NETWORKS

In single connected network research, the majority of the algorithms follow centralised architecture. For example, in [68] a centralised packet scheduling algorithm for single connected networks is proposed to enhance the performance of URLLC for multi-cell systems. To reduce the complexity the paper proposes a sub-optimal algorithm. The low complexity algorithm reduces the packet delay in the queue that ensures the reliability and latency requirements of URLLC. A similar work could be found in [69] for radio resource allocation in multi-cellular networks in 5G systems. The paper shows 90% more latency reduction than the conventional distributed scheduling algorithms. In dense cellular networks, UEs belong to the communication range of more than one base station. In this case cell association is an important factor that affects the performance of the network. Considering the fact [70] proposes a centralised joint cell selection and scheduling algorithm. The low complexity algorithm reduces the queuing delay. The proposed algorithm is evaluated for multi-cell and multi-user environments and shows the significant improvement of the performance over the distributed solutions. The above methods allocate the whole PRBs among the UEs through a centralised decision. However, full allocation of PRB leads to inefficient use of bandwidth [71].

Packet scheduling for Integrated Access and Backhaul (IAB) networks is proposed in [72]. To reduce the optical fibre link for each base station the 3rd Generation Partnership Project (3GPP) introduces the IAB. In the integrated policy the access to the UEs and the backhaul links share the same hardware, protocol stack and the spectrum. For the integrated system a semi-centralised scheduling algorithm is proposed for resource allocation. The low complexity algorithm shows the better performance in terms of throughput and E2E delay.

A cooperative scheduling is proposed in [73] for Cloud Radio Access Network (C-RAN) to maximise integrated system throughput. The intra-base station cooperation based scheduling method is used to schedule the URLLC, eMBB and mMTC traffic of 5G networks. A joint optimisation of time/frequency-domain scheduling and frequency allocation is used to maximise the integrated system throughput. The joint optimisation increases the throughput of the cell-edge terminals and thus increases the fairness among the UEs. However, the detailed analysis of packet loss ratio and latency are required to find the applicability of the method. Moreover, the complexity of the method is high.

2) SCHEDULING IN MULTI CONNECTED NETWORKS

Multi connected network is a promising architecture to increase the reliability and capacity of wireless network [74]. It provides flexibility by transmitting multiple copies of the same information through multiple links from source to destination [75]. A number of works could be found in literature that consider centralised scheduling in multi-connected networks. For example, in [76] the authors propose Split Responsibility Scheduler (SRS) for 5G networks. The functionalities of the scheduler are divided into two parts: namely Resource Regulator (REG) and Resource Allocator (RA). REG is responsible for handling the service behaviour and RA is responsible for controlling the radio behaviour. In traditional scheduler, both functionalities are handled by the scheduler [77]. Different vendors use different strategies to implement the functionality. Adding new functionalities to the existing scheduler is complicated for the vendors since

TABLE 6. Key features of optimisation based approach.

Reference No.	Features	Downlink	Uplink	Single cell	Multi cell	Multi channel	No. of user
[47]	• Multiple unreliable transmissions are combined. • Multiple transmission based optimisation. • Knapsack-inspired heuristic technique.	✓		✓		✓	Variable 2 - 14
[50]	• Deals with ultra reliability and very low latency without perfect SCI. • Same packet is transmitted simultaneously over multiple channels exploring frequency diversity. • The feedback from the UE is used for optimal decision.	✓		✓		✓	Variable 4 - 5
[54]	• Joint admission control and resource scheduling. • Standard continuous and binary SNR models. • Low complexity approximation algorithms.	✓		✓		✓	Variable 2 - 100
[55]	• Real-time multi-cell and multi-channel URLLC scheduling. • Scheduling algorithm with a feasible set for schedulability.	✓			✓	✓	Variable 90
[58]	• QoS differential scheduling. • FIFO service discipline with Markov chain. • Similar states aggregation and low complexity solution.	✓		✓		✓	Not specified
[62]	• Jointly optimisation of PRB and power allocation. • Convex problem derivation with a approximation algorithm.	✓		✓		✓	Variable 2 - 9
[63]	• Priority based uplink resource scheduling. • Dynamic scheduling is adopted for QoS. • Balances system throughput and fairness.		✓	✓			Not specified
[64]	• uplink resource allocation for heterogeneous cellular networks. • Optimisation technique to maximise the system throughput. • Tailor expansion based solution.		✓		✓	✓	Variable 5 - 20

the testing and tuning the new functionality is not possible with the existing system. Thus the proposed SRS solution is suitable for current and future multi-connected networks. An important feature of the reorganisation (i.e., dividing into REG and RA) is that the architecture can be configured to run both centralised and distributed algorithms. To this end, the paper implements both centralised and distributed algorithms. The method is evaluated with a simple version of centralised and distributed algorithm. However, a more realistic environment is required to verify the effectiveness of the SRS.

To increase the flexibility in interaction of 5G network with the legacy LTE in heterogeneous environment [78] proposes a resource allocation algorithm that supports a wide range of service requirements. The UEs of 5G are expected to connect both networks simultaneously. In order to allocate the resources an optimisation approach is used that maximises the throughput in the system. Then a centralised solution is proposed. To overcome the limitations of the centralised algorithm, authors in [79] proposed a distributed solution. The distributed algorithm demonstrates the performance similar to the centralised algorithm when the number of UEs are less than or equal to 20. However, the distributed RRM achieves the energy, spectral and processing cost efficiency, but the paper does not consider isolation and service orientation in the model development [80].

A resource allocation scheme is proposed in [81] for Coordinated Multipoint (CoMP) enabled URLLC with centralised

C-RAN architecture. The main challenges of the environment include fronthaul capacity and remote radio head resource availability constraints. To mitigate the challenges the paper proposes a packet delivery mechanism, queuing strategy and time-frequency resource allocation for URLLC in the C-RAN architecture. An optimisation technique is used to minimise the total allocated bandwidth. A heuristic algorithm is used to allocate optimum resources among the UEs. However, the algorithm is designed only for the URLLC traffic without considering the eMBB and mMTC traffic. Moreover, the issue of computational overhead is ignored in the paper. The key features of single and multi-connected scheduling algorithms are summarised in Table 7.

V. JOINT URLLC AND EMBB PACKET SCHEDULING ALGORITHMS

The packet scheduling algorithms described in the above section are mainly focused on the URLLC scheduling algorithms. However, the traffic of 5G networks consists of three different types of services including URLLC, eMBB and mMTC. Therefore, there are many algorithms that jointly schedule these different types of traffic together. This section provides a thorough overview of the joint scheduling techniques for 5G networks which also includes ML based techniques.

Optimisation based approach is one of the most preferable techniques for researchers to schedule the traffic in 5G networks. In a 5G network URLLC packets may arrive to the scheduler during the ongoing eMBB transmission. In this case, how to schedule the URLLC packets is challenging. In the literature two different optimisation techniques are adopted including dynamic and semi-persistent approaches. Dynamic scheduling can be classified into puncturing scheduling and preemptive scheduling [82]. In scheduling with puncturing technique, the packets are prioritised and dynamically overlapped at the mini slot boundary when URLLC packets are arrived during ongoing eMBB transmission. In preemptive scheduling the resources are reserved preemptively for URLLC packets before the packets actually arrive in the system [83]. On the other hand URLLC packets are pre-scheduled in semi-persistent technique with a fraction of bandwidth [84], [85]. Fig. 3 and Fig. 4 show a typical preemptive and a typical semi-persistent scheduling techniques respectively.

A. DYNAMIC SCHEDULING

A joint scheduling of URLLC and eMBB traffic is proposed in [86]. In this technique, time is divided into slots with 1 ms duration. Each slot is further divided into 8 mini slots. The eMBB traffic is scheduled at slot boundaries. Due to time latency requirement, the URLLC traffic is scheduled immediately at mini slot boundaries. A joint scheduling problem is formulated as a joint optimisation problem. In the scheduling maximum resources are allocated for eMBB with minimum amount is allocated to URLLC which immediately satisfies the URLLC demand. This immediate insertion of URLLC traffic in mini slot (within a slot of eMBB) boundaries is defined as superposition/puncturing. The problem is solved with three different models of superposition/puncturing the eMBB slot. The three orthogonal access mechanisms provide interference-free mini slot allocation for URLLC traffic but it degrades the eMBB spectral efficiency as the total number available symbols for transmission is decreased [87] due to the superposition/puncturing.

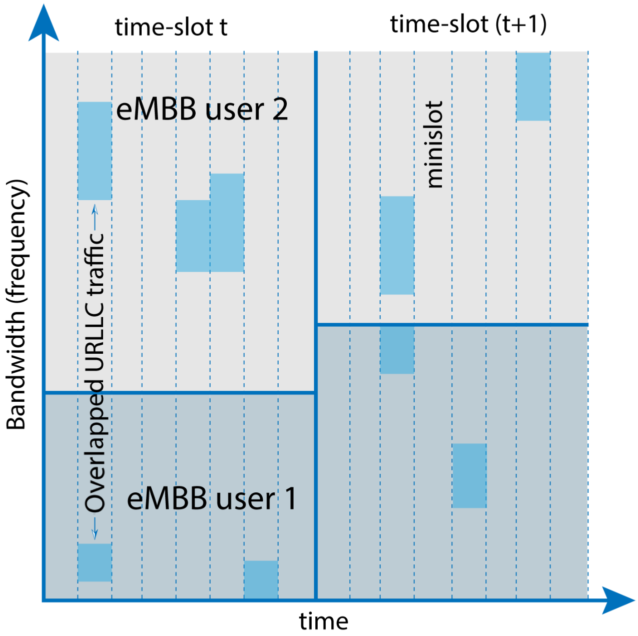


FIGURE 3. An illustration of a preemptive scheduling scheme in 5G.

A user–centric joint URLLC and eMBB scheduling algorithm is proposed in [88]. A Null-Space-Based Preemptive Scheduler (NSBPS) is introduced in the paper for densely populated 5G networks. A cross-objective optimisation is used for the scheduler that fulfils the stringent QoS requirements of URLLC while minimising the loss in eMBB traffic. System spatial degrees of freedom is used to find the interference-free space for the critical URLLC traffic. However, in the algorithm, if no URLLC traffic exists then the eMBB traffic is scheduled with the PF algorithm. When the URLLC packets are arrived, the weighted PF is used to schedule the URLLC according to following relation:

$$S_{i,t}^{NSBPS} = \arg\max_{i} \frac{r_i(t)}{R_i(t)} \beta_{k_{URLLC}}, \tag{17}$$

where $\beta_{k_{URLLC}}$ is the scheduling constant for URLLC traffic.

A similar work of user-centric joint URLLC and eMBB scheduling algorithm can be found in [89]. The joint Multi-User Preemptive Scheduler (MUPS) algorithm finds suitable eMBBs where the URLLC traffic can be preempted. However, the scheduler instantly transmits the URLLC packet suspending the ongoing eMBB transmission which is the major obstacle to meeting the interruption time of 5G NR [90]. It is expected that 5G NR would support 1080p, 2K, 4K, 8K full

TABLE 7. A summary of scheduling techniques in single and multi-connected networks.

Category	Reference No.	Features	Traffic			Complexity
			URLLC	eMBB	mMTC	
Single connected	[68]	• Multi-cell scheduling algorithms with a centralised scheduling that enhances URLLC performance • It reduces the queuing delay of URLLC packets	✓			Medium
	[69]	• Centralised packet scheduling that supports the QoS requirements of URLLC • Both analytical and simulation results are evaluated	✓			Medium
	[70]	• Joint cell selection and scheduling algorithm for URLLC • A realistic multi-cell, multi-user dynamic network is used in the performance study	✓			Medium
	[72]	• Semi-centralised resource allocation scheme for IAB network • Flexible and low complexity algorithm		✓		Low
	[73]	• Intra-BS cooperation in C-RAN heterogeneous network • Joint optimisation of time/frequency-domain scheduling and bandwidth allocation to maximise system throughput	✓	✓	✓	High
Multi connected	[76]	• Scheduling functionality is divided into REG and RA • Both centralised and decentralised algorithms can be implemented with the proposed system • Adding and testing new scheduling behavior is easy with the SRS	✓			Medium
	[78]	• Optimisation based resource allocation algorithm for LTE, 5G multi-connectivity networks • Both centralised and decentralised algorithms are developed • Both algorithms show similar performance with lower traffic load	✓			High
	[81]	• CoMP enabled URLLC for C-RAN • Heuristic technique is applied to allocate the resource among the UEs	✓			High

HD video resolution with less than 1 ms mobility interruption time, which is likely to be challenging with this technique in high mobility condition.

In [91] a risk sensitive based resource allocation technique is proposed for URLLC and eMBB traffics that ensures the requirements of URLLC. An optimisation problem is formulated to minimise the risk of eMBB traffic. In contrast to traditional average-based formulation, the risk sensitive based formulation is introduced with a conditional value of risk that reduces the eMBB loss rate while ensuring the latency of URLLC. The method punctures the high data rate eMBB users with high probability in good channel condition while protects the low data rate eMBB users in bad channel condition.

A dynamic multi-connectivity based joint scheduling algorithm with traffic steering for URLLC and eMBB traffic is proposed in [92]. The framework slices the URLLC and eMBB each other to avoid the URLLC packet queue. An optimisation problem is formulated for joint resource allocation. A modified Effective Capacity (EC) is used to evaluate the performance of the framework. A two-step solution is proposed considering queue length and the EC model. The multi-connectivity based model increases the reliability and the number of packets in the network for densely populated

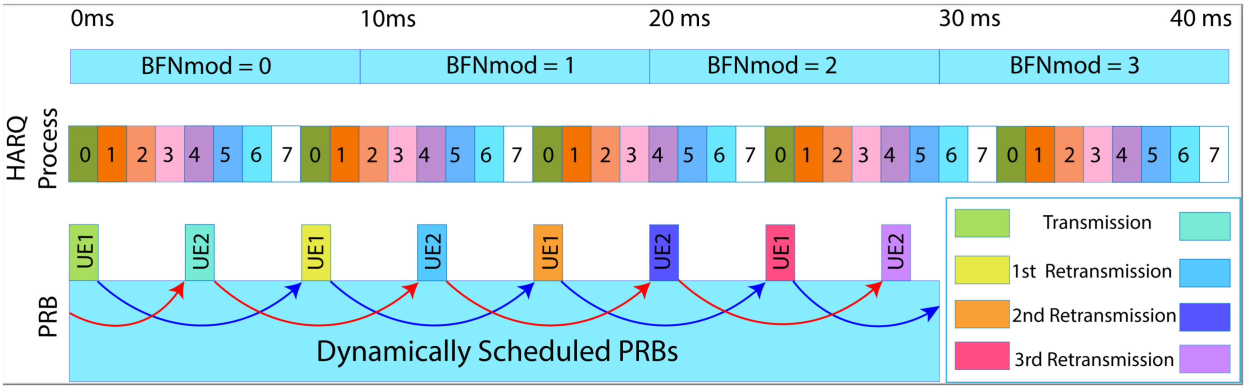


FIGURE 4. An illustration of typical semi-persistent scheduling.

networks. At the same time it increases the eMBB packet loss rate [93].

A novel latency, control channel, HARQ and radio channel aware joint packet scheduling algorithm is proposed in [94]. The proposed algorithm avoids costly segmentation of URLLC payloads over multiple transmissions and employs the gain of frequency-selective multiuser scheduling. The scheduler uses two different metrics to schedule URLLC and eMBB traffic. Suppose $r_i(t)$ denotes the throughput achieved by the scheduler at time t to user i. Then a URLLC packet is scheduled that maximises the following metric:

$$S_{i,t}^{URLLC} = \arg\max_i \frac{r_i(t)}{r_i(t)}, \tag{18}$$

where $r_i(t)$ is the instantaneous full-bandwidth throughput, and eMBB packet is scheduled according to the PF metric as follows

$$S_{i,t}^{eMBB} = \arg\max_i \frac{r_i(t)}{R_i(t)}. \tag{19}$$

The low complexity algorithm shows latency improvement in URLLC packets and higher eMBB throughput. However, the method satisfies the QoS requirements of URLLC but cannot guarantee sufficient isolation requirement among different service slices under the high URLLC traffic load and hence eMBB users may not get enough RBs [95].

An optimal resource allocation for aperiodic URLLC traffic is proposed in [96]. The technique maximises the URLLC admission with minimum preemption of the eMBB traffic. The joint optimisation technique finds a policy of which eMBBs are suitable for preemption and how many resource blocks are allocated in the preempted packets. Two different solutions are proposed in the paper that finds a close-to-optimum solution.

In [97] a dynamic joint scheduling algorithm is proposed to schedule the URLLC and eMBB traffic at frame level. Real time queue monitoring is used to ensure the latency of URLLC traffic. A complicated URLLC packet outage probability (i.e., packet drop probability) is analytically derived. Then a stochastic optimisation problem is formulated that maximises the eMBB throughput while maintaining the constraints of URLLC. A low complexity two stages solution is proposed in the paper.

A joint scheduling algorithm is proposed in [98] with minimum impact on eMBB traffic. The preemptive scheduler exploits a precoder compatibility estimate with a metric of degree of similarity between different Multiple-Input Multiple-Output (MIMO) channels. The metric is used to find which of the ongoing eMBB transmission will be paused without recomputing the precoding matrix. This reduces the MIMO precoder computation overhead and URLLC demodulation pilot overhead. The method increases the URLLC transmission performance and the eMBB transmission rate. But the performance degrades when the URLLC traffic increases significantly as more eMBB packets are punctured.

A low complexity near optimal radio resources scheduling is proposed in [82] to minimise the conflict between the URLLC and eMBB traffic. In conventional technique the eMBB throughput is maximised while maintaining the URLLC requirements. To minimise the conflict an optimisation problem is reformulated. Then to find the near optimal solution conflict-aware heuristic is proposed. The lightweight heuristic algorithm is a family of greedy algorithms which is designed motivated by the bin packing problem optimisation. However, the lightweight heuristic degrades the performance especially for dense networks as more eMBB packets are punctured.

A punctured scheduling algorithm is proposed in [8] without prior reservation of transmission resources for URLLC traffic. In the technique when a URLLC packet arrives in the system it immediately schedules the packet by puncturing the ongoing eMBB transmission. An optimal multiplexing of the URLLC and the eMBB traffic is used in the paper to minimise eMBB loss rate with low modulation and coding scheme index. A recovery mechanisms for the eMBB transmission is used to minimise the eMBB packet loss with puncturing-aware dynamic link adaptation and eMBB-aware scheduling decisions.

A joint metric based resource allocation technique is proposed in [99] to minimise the loss of eMBB with optimal URLLC placement. The metric is used to allocate resources for URLLC traffic at any mini-slot boundary when required. The throughput to average metric is used to schedule the URLLC traffic and PF metric is used to schedule the eMBB traffic.

A superposition/puncturing based resource allocation problem is formulated in [100] for joint resource and power allocation in URLLC and eMBB traffic for 6G networks. The problem is Mixed Integer Nonlinear Programming (MINLP) which is NP-hard. To solve the problem, a low complexity algorithm is proposed that achieves higher reliability and higher eMBB data rate compared to existing algorithms.

A fronthaul network scheduling architecture and scheme is proposed in [101]. The method is based on Burst Limiting Shaper (BLS) mechanism to provide QoS guarantees to heterogeneous traffic including URLLC and non-URLLC. The paper uses the BLS reserved capacity to schedule the low priority traffic dynamically. This technique increases the overall throughput and reduces the latency.

B. SEMI-PERSISTENT SCHEDULING

In [102], a dynamic resource allocation algorithm is proposed for IoT services. The authors use a dynamic optimisation model for scheduling URLLC and eMBB services with RAN slicing. The model is based on a cost function that considers power consumption and service quality in both time and frequency domain. However the authors did not consider the multi-tenant multi-tire in their formulation which is most likely in real world applications [103].

Different resource allocation schemes for transmission and re-transmission for industrial IoT are proposed in [104]. The paper investigates the individual resource reservation scheme versus a pool of contention-based reservation schemes. Then results are used to propose a novel resource allocation scheme for packet transmission and re-transmission and drive corresponding analytical models for packet loss. The paper then finds an optimal parameter setup that allows meeting the URLLC requirements with low resource consumption. However, for high rate of deterministic or sporadic traffic the replicas increase the eMBB loss rate [105].

A low complexity and efficient downlink semi-persistent scheduling algorithm is proposed in [84] based on adaptive short term traffic prediction. The novel scheduler achieves high throughput, fairness and low latency. The algorithm pre allocates radio resources for mobile users based on short term prediction of arriving traffic over multiple TTIs. Two functions are defined in the paper considering instantaneous channel conditions, historical data rates, the buffer occupancy and the predicted traffic. These two functions are used to schedule the packets by assigning the scheduling priority to each user.

In most of the joint scheduling approaches all the resources are allocated to the eMBB users at first. Whenever a URLLC service request arrives, the system instantly responds to the request by puncturing the eMBB traffic in the next mini-slot. The method does not consider reasonably the impact of the puncturing that reduces the data rate of eMBB [106]. Considering the fact the paper proposes flexible resource allocation by optimising the resource allocation and puncturing weight matrix with Block Coordinate Descent (BCD) algorithm so that the URLLC devices can capture resources with minimum data rate loss of eMBB. However, The summary of the major joint scheduling algorithms are presented in Table 8.

C. ML BASED JOINT SCHEDULING

Most of the optimisation techniques discussed in the previous subsection do not have a closed-form solution. Either the objective function or the constraints are complicated and hence suboptimal solutions are evaluated using the model based optimisation techniques. To this end, the machine learning technique provides a good solution to learn the optimal algorithm. In the literature different machine learning based approaches could be found. For example, a model-free Deep Reinforcement Learning (DRL) based solution, DEMUX is proposed in [107] for joint URLLC and eMBB scheduling that minimises the adverse impact of puncturing scheduling. The DEMUX is a deep function approximator that determines the preemption solution in each eMBB TTI. The technique ensures fast and stable convergence of learning the neural network by exploiting the intrinsic property of the problem and obtains the scheduling of the URLLC and eMBB traffic while minimising the data loss of eMBB.

Machine learning based Self-adaptive Flexible TTI Scheduling (SAFE-TS) is proposed for joint URLLC and eMBB scheduling in [108]. Random Forest based Ensemble TTI Decision Algorithm (RF-ETDA) is used to compute the TTI for each service. The proposed method improves the performance for URLLC services compared to the other machine learning based methods.

A coexistence scenario of URLLC and eMBB traffic over 5G NR is addressed in [109] where the QoS demand of both URLLC and eMBB traffic are simultaneously important. For the scenario an AI-enabled reinforcement learning-based algorithm is proposed to improve the efficiency of both URLLC and eMBB users. The algorithm optimises the latency, reliability of URLLC and throughput of eMBB users. To achieve the optimisation the paper proposes a multi-agent Q-learning based joint power and resource allocation.

CSI reflects the channel condition among the base stations and UEs. Accurate CSI affects the performance of the transmission. Considering the importance [110] proposes a Deep Learning (DL) based CSI estimation technique for highly mobile vehicular networks. To this end, the paper formulates and solves dynamic network slicing based URLLC and eMBB scheduling problems for the Vehicular User Equipments (VUEs) of the networks. The problem is formulated in a way that minimises the threshold of rate violation probability for eMBB slice while maintaining a

TABLE 8. Summary of joint scheduling algorithms.

Reference No.	Main consideration	Target applications	uplink /downlink	Scheduler type	Duration Slot	Mini-slot	Closed-form /Approximation
[86]	Dual objectives of minimising eMBB loss rate and immediate scheduling of URLLC traffic	All applications of URLLC and eMBB	Downlink	Puncturing	1 ms	0.125 ms	Closed-form /Approximation [1]
[88]	Cross-objective optimisation that maximise the eMBB ergodic capacity while maintaining the requirement of URLLC	All applications of URLLC and eMBB	Downlink	Preemptive	1 ms	0.143 ms	Approximation
[89]	MUPS Cross-optimises network performances to achieve maximum spectral efficiency and URLLC latency	All applications of URLLC and eMBB	Downlink	Preemptive	1 ms	0.143 ms	Approximation
[91]	Risk sensitive based optimisation problem is formulised to reduce the eMBB loss rate	All applications of URLLC and eMBB	Downlink	Puncturing	1 ms	0.125 ms	Approximation
[92]	Dynamic multiconnectivity based joint scheduling with network slicing	All applications of URLLC and eMBB	Downlink	Preemptive	1 ms	0.125 ms	Approximation
[94]	low complexity payload and control channel aware joint scheduling	All applications of URLLC and eMBB	Downlink	Preemptive	1 ms	0.143 ms	Approximation
[96]	Maximising the URLLC transmission while minimise the data loss rate of eMBB traffic.	All applications of URLLC and eMBB	Downlink	Preemptive	1 ms	0.143 ms	Approximation
[97]	Dynamic joint scheduling at frame-level through stochastic optimisation with queuing mechanism exploring the packet drop probability	All applications of URLLC and eMBB	Downlink	Preemptive	1 ms	0.125 ms	Approximation
[98]	Similarity between different MIMO channels metric based joint scheduling algorithm to reduce eMBB traffic loss rate that reduces the computation overhead	All applications of URLLC and eMBB	Downlink	Preemptive	1 ms	0.071 or 0.143 ms	Approximation
[82]	Reformulation of maximising eMBB throughput as minimising the conflicts among the URLLC and eMBB traffic	All applications of URLLC and eMBB	Downlink	Puncturing	1 ms	0.125 ms	Approximation
[8]	optimal multiplexing the URLLC and eMBB traffic with low modulation and coding scheme index	All applications of URLLC and eMBB	Downlink	Puncturing	1 ms	0.143 ms	Approximation
[99]	Optimised metric based URLLC and eMBB scheduling	All applications of URLLC and eMBB	Downlink	Puncturing	1 ms	0.143 ms	Approximation
[102]	Power consumption and service quality based dynamic optimisation technique for RAN slicing and scheduling	IoT	uplink	Semi-persistent	0.5 ms	Variable	Approximation
[104]	For deterministic traffic pool based reservation and for sporadic traffic a contention-based scheme is proposed with replicas.	Industrial IoT	uplink	Semi-persistent	1 ms	0.144 ms	Approximation

minimum probabilistic threshold rate criteria for URLLC slice. The approach significantly reduces CSI overhead of eMBB vehicles.

A DRL based network slicing algorithm is proposed in [111] to slice the total available resources between the URLLC and eMBB traffics. In the algorithm the full time-frequency resource is allocated for eMBB then an optimal policy is trained dynamically to allocate the resource by puncturing the eMBB codewords. With the assumption of a limited amount of puncturing the eMBB traffic is tolerable, the paper shows that the policy never violates the latency and reliability of URLLC while maintaining a minimum rate loss of eMBB traffic.

In [112] the resource allocation problem is converted to a Markov Decision Process (MDP) problem considering time-varying channels between mobile station and the base station. To solve the optimisation problem, a Q-learning algorithm is introduced. Since the state space in the learning is enormous, the paper uses Bellman equation and the Deep Q-network (DQN) based resource allocation algorithm.

A resource slicing problem is formulated to allocate resources to URLLC and eMBB users in [113]. The main goal of the paper is to maximise the eMBB data rate subject to satisfy the reliability and latency constraints of URLLC. An optimisation-aided DRL is proposed to solve the problem. The method combines the advantages of both optimisation and machine learning techniques for efficient resource allocation. In the optimisation phase, the problem is decomposed into several convex subproblems to obtain a solution for each of the subproblems. In the learning phase DRL algorithm is proposed to intelligently distribute the URLLC traffic among the eMBB traffic.

To schedule URLLC and eMBB traffics [114] proposes to achieve QoS tradeoff between the traffics in 5G networks. The paper jointly optimises the bandwidth allocation and overlapping position of URLLC traffic in the eMBB traffic. A DRL method based on deep deterministic policy gradient with prioritised sampling is applied to learn the tradeoff. The DRL achieves the long-term QoS tradeoff with the observation of environment variables. However, the summary of the major machine learning based joint scheduling algorithms are presented in Table 9.

VI. PERFORMANCE COMPARISON OF SOME SELECTED SCHEDULING ALGORITHMS

A. EXPERIMENTAL SETUP

To evaluate the performance we have simulated a network with NS-3 simulator software V3.28.1. The considered network is a rectangular area of 4 Km X 2 Km. In the topology the UEs are deployed randomly in the rectangular area. Two Evolved NodeBs (eNB) are located at the location (1.0 Km, 1.0 Km) and (3.0 Km, 1.0 Km). Each eNB is connected to a Packet data network Gateway (PGW) with default Evolved Packet Core (EPC) point to point link. The network contains two remote servers. The servers generate the data packets for UEs. The packets are sent to the UEs through the PGW. The data rate between the remote host and the PGW is 100 Gbps and the delay is 0.01 ms. The remote hosts generate the stream of data packets with specific random time intervals and transmit to PGW. The PGW transmits the received packet from the remote host to the eNB. The eNB performs the radio resource scheduling and other functions to transmit packets to UEs. We use the LTE framework and change the parameters similar to [115]. We use $k = 2$ OFDM symbols, subscriber spacing is 15 kHz and OFDM symbol duration is 71 ms.

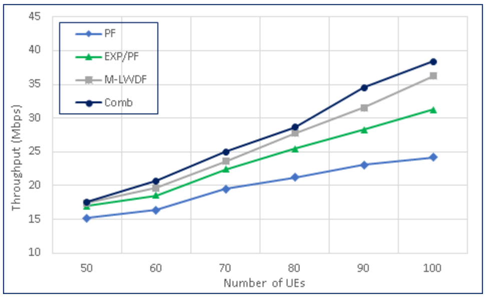


FIGURE 5. System throughput over number of UEs.

We have simulated the network for streaming applications. The bandwidth of the system is 5 MHz. We ran the simulation for 60 seconds. The simulation parameters are listed in Table 10. In the following subsection we have briefly described the network topology and simulation parameters. Then we have analysed the comparative results of the chosen algorithms including PE, EXP/PF, M-LWDF and the combined scheduling algorithms.

B. SIMULATION RESULTS

In order to evaluate the efficiency of the considered scheduling algorithms, the different performance indicators such as throughput and delivery ratio are compared. We have used the number of UEs from 50 to 100 in the interval of 10. Packet generation time for UEs is random. This means the remote hosts generate packets with different rates for different UEs. The average packet inter generation time for different UEs are given in Table 11.

1) SYSTEM THROUGHPUT

The total system throughput is the sum of all user throughput within the cell. The total system throughput for different numbers of UEs is shown in Fig. 5. The result shows the throughput of PF, EXP/PF, M-LWDF and the proposed scheduling algorithm for the same network setup in each case. From the figure we see that the combined scheduling algorithm shows the highest throughput whereas the PF algorithm achieves the lowest throughput compared to all other simulated algorithms. The combined scheduling algorithm and M-LWDF achieve nearly similar throughput for a varying number of UEs. The closest performance of PF is shown with the EXP/PF algorithm. The figure also shows the performance comparison between PF and other algorithms.

2) PACKET DELIVERY RATIO

At the beginning of simulation a distinct packet generation rate is assigned to the remote hosts to generate packets for each UE. Thus we compare the packet delivery ratio without comparing the fairness index. The packet delivery ratio is defined as the ratio of the number of packets received by the receiver to the number of packets transmitted by the transmitter to the receiver. We compute the average packet delivery ratio of all the UEs for each simulation. The average

TABLE 9. Summary of machine learning based approach.

Reference No.	Key features	Learning method
[107]	• Model-free DRL based solution for joint scheduling • Deep function approximator • Fast and stable convergence of learning the network	DRL
[108]	• Machine learning based flexible TTI selection • The selected TTI is used for resource allocation and scheduling	RF
[109]	• Q-learning based joint power and resource allocation algorithm • Improves the reliability and latency of URLLC and data rate of eMBB	RL
[110]	• DL based CIS estimation technique • Dynamic network slicing based resource allocation problem for vehicular UEs with the CSI	DL
[111]	• DL based network slicing • Full resources are allocated for eMBB traffic then it is punctured for URLLC traffic	DRL
[112]	• The resource allocation problem is converted to MDP problem • Bellman and DQN based resource allocation	DRL
[113]	• Optimisation-aided DRL • Optimisation technique is used to solve the resource allocation • DRL is used to distribute the URLLC traffic among eMBB traffic	DRL
[114]	• Jointly optimisation the bandwidth allocation and the overlapping position of URLLC and eMBB traffic • Long-term QoS trade-off is achieved with the DRL	DRL

TABLE 10. Simulation parameters.

Parameters name	Values
Application type	Streaming application
Packet generation interval	Random
Simulation time	20 seconds
Bandwidth	20 MHz
AMC mode	Piro
Cell radius	1.5 km
Transmission power of eNB	40 dB
Transmission power of UE	30 dB

TABLE 11. Average packet inter generation time of different UEs.

Number of UEs	Average inter generation time (ms)
50	66.12
60	52.42
70	51.75
80	53.28
90	53.54
100	51.43

packet delivery ratio for different numbers of UEs is shown in Fig. 6.

The figure shows that the average packet delivery ratios of EXP/PF, M-LWDF and the combined scheduling algorithms are very close over the number of UEs from 50 to 100 compared to the packet delivery ratio of PF. The combined and PF scheduling algorithms show the highest and lowest delivery ratio, respectively. Although the delivery ratios of EXP/PF, M-LWDF and the combined scheduling algorithms are close to each other, the M-LWDF is the closest to the combined scheduling algorithm. The difference

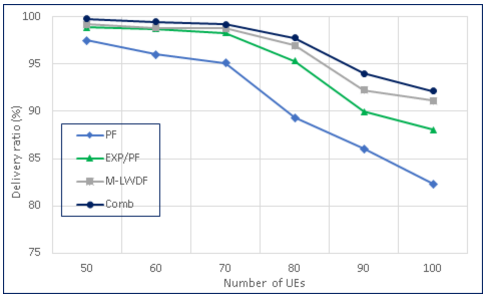


FIGURE 6. Average packet delivery ratio over number of UEs.

between the two is slightly noticeable when the number of UEs are more than 80.

VII. CHALLENGES FOR SCHEDULING IN EMERGING 6G SYSTEMS

Despite the success of 5G networks, the real world deployment exposes the limitations of the existing 5G system driven by the ever increasing demand of data rate, new applications and widespread penetration of wireless connectivity in all spheres of our everyday life [2]. It is expected that the data rate of 6G will be up to 1 Tbps. Moreover, the latency bound of 6G is at least ten times higher than 5G. Applications such as autonomous and connected driving [116], Vehicle to Everything (V2X) communications [117] and massive IoT enabled industry automation [118] would clearly require much more enhanced latency and reliability requirement along with other generic system requirement. Hence the algorithms which are suitable for 6G needs to be designed.

Terahertz (THz) communications is expected to be one of the key enabling technologies for the 6G network. Due to the nature of THz communication technologies the design of 6G causes some MAC layer design challenges including deafness problem, coupling problem and transmission scheduling problems. Specially the scheduling problem is challenging in the heterogeneous service environments with the low computational capacity UEs. To further enhance latency performance more innovative solutions for scheduling needs to be designed.

Various ML algorithms will also play a crucial role in future 6G systems. The work in [119] explores various challenges in resource scheduling for next generation networks and potential role of ML variants in solving them. It developed a delay aware traffic scheduling algorithm which employs DRL and was able to achieve lowest average delay compared to state of the art benchmarks. However, significant challenges are still there. For example, adversaries can employ ML algorithms to breach security by intelligently jamming the signal leading to failure of the scheduler performance. Thus further exploration is required for successfully deploying machine learning. Emerging techniques such as federated learning can also be employed to further enhance the performance.

A. LATENCY

The current 5G systems can achieve high reliability with the cost of high latency from ten to hundred milliseconds. Thus the system cannot achieve the 1 ms latency requirement of URLLC with its basic components including OFDM numerology, radio frame structure, HARQ, Modulation and Coding Schemes (MCSs) etc. [6]. To achieve the latency requirement, even for a single flow the basic components of the 5G need to be modified.

To achieve the requirements of multiple flows upper layer optimisation is required [120]. Specifically radio resources should be scheduled considering the stringent latency requirement. This is challenging in 5G since huge bandwidth of the network is allocated for eMBB traffic. Moreover, the huge amount of mMTC devices increases extra difficulties in the scheduling algorithm which consumes a significant portion of the resources of the network.

B. RELIABILITY

Reliability of transmission depends on multiple factors including MCSs, time/frequency resources, number of antennas and transmit power [121]. 5G uses an MCSs palette that has a capacity boundary close to the Shannon limit. In 5G, base stations transmit with a MCS which provides guaranteed delivery of a transport block with 90% probability. To increase the reliability additional transmission is required. For example to increase the reliability by ten times, nine additional retransmissions are required which is impossible within the delay bound. A possible solution of the problem is to reduce the data transmission time with short packet size that requires less data processing and checksum computing time. Another solution is to use advanced decoding and early-feedback technique where time is divided into mini-slot with 0.125 ms duration and use additional retransmission within the delay bound limit to achieve the reliability. The Ultra Mini Slot Transmission (UMST) has been investigated for future 6G systems [122] which offers a promising improvement in terms of latency performance compared to existing 5G systems. Much more work is required to design a system that fulfils the 6G network requirements.

C. FLEXIBLE TECHNOLOGY DEVELOPMENT

Flexible technology concerns the ability to extend the technology so that it can be adapted to changing circumstances [123]. The LTE network has limited flexibility because it does not satisfy the requirements of 5G or 6G. For example, the Transmission Time Interval (TTI) of LTE is 1 ms and each frame consists of 10 TTIs. As a result the networks cannot satisfy the latency requirement of URLLC. To satisfy the latency requirement short frame structure is proposed in [124]. Similarly, the existing LTE scheduling algorithms including PF, EXP/PF, LOG rule, EXP rule and LWDF cannot be used for URLLC in 5G networks. The scheduling algorithms do not consider the latency bound of the URLLC. Thus flexible technology is required to adapt with the changing requirements for sustainable development.

D. HETEROGENEOUS SERVICES

The 5G network is designed targeting three services including eMBB, URLLC and mMTC. At present the 5G network is establishing targeting mostly the eMBB service. In the literature we found a number of uplink and downlink URLLC packet scheduling algorithms and joint scheduling algorithms of eMBB and URLLC. On the other hand some works are found on mMTC uplink scheduling algorithm for example, [125], [126], and [127]. But few works could be found on downlink joint eMBB, URLLC and mMTC scheduling algorithms. mMTC devices mostly upload data of environmental phenomena, thus it is ignored in the downlink scheduling algorithm. In future 6G systems, downlink scheduling is expected to be an essential feature to send various control and information to mMTC devices.

E. SCHEDULING IN MULTI CONNECTED NETWORKS

An important way to increase the reliability of the network is multi-connectivity, in which users are enabled to connect multiple base stations simultaneously [128] The multi-connectivity solution also achieves high mobility and load balancing. However, to achieve the QoS of scheduling algorithms in multi-connected networks we need to consider the radio interface of different cells in addition to traditional time, frequency, space and power consideration of a single cell. In the existing literature very few works could be found on scheduling in the multi-connected networks. Moreover, ML based scheduling algorithm is one of the future research directions for 6G wireless network that requires further investigation.

F. RESOURCE SCHEDULING IN CLOUD ENVIRONMENT

Resource scheduling in cloud environments decreases the execution and computation time of cloud workloads with reduced packet deadline violation time [129]. Resource scheduling in a cloud environment is a challenging job due to heterogeneity and uncertainty. The scheduling problem in the cloud environment cannot be solved with existing scheduling algorithms. Thus cloud-oriented URLLC applications need more attention to achieve the QoS in the cloud environment. There are several issues including energy management, data security and dynamic scalability in the cloud environment that needs more attention for investigation.

G. PUNCTURING THE EMBB PACKET

In 5G and beyond systems an eNB node performs scheduling. If an urgent packet arrives at eNB it cannot schedule the packet until the next slot boundary. But this may violate the stringent delay bound of URLLC. To solve the problem the time slot is divided into mini-slot (usually 0.125 ms duration). When whole resources are occupied by eMBB transmission and an URLLC packet is arrived then the eNB punctures the ongoing eMBB transmission by transmitting the URLLC packet in the next mini-slot duration. This causes some losses of resources and the punctured eMBB user recovers the loss using retransmission. The frequent preemption degrades the performance of the network. Scheduling URLLC packets minimising the impacts of eMBB packets is challenging in a heterogeneous service environment.

H. PROCESSING TIME

Reducing processing time is important in downlink transmission for 5G and the 6G systems. Particularly the downlink retransmission within the stringent delay bound is challenging, because when the transmission is failed the packet reaches close to the delay bound limit. However, a significant amount of time is spent by UE for processing. For example, a typical 5G receiver spends 60% of the processing time for turbo decoding and the remaining time is spent for other operations including OFDM processing and other operations [130]. Hence low complexity decoding and other processing algorithms is important for the UEs devices.

ML technique and its variants will play a crucial role in this regard. The work in [131] proposed a joint scheduling and resource slicing mechanism for 6G URLLC systems with particular focus to vehicular network. Federated learning algorithm developed in the work has demonstrated significant performance improvement. It has been demonstrated earlier that using ML for channel estimation for both conventional wireless system and vehicular network helps to reduce complexity and thus processing time [132], [133]. Several works also demonstrated promising performance in emerging 6G systems by incorporating various state of the art techniques such as edge Artificial Intelligence (AI) [134]. However, it is evident from the works and requirements that much more investigation is needed to realise the full potential of the AI/ML techniques.

A joint time-slot scheduling sub-band scheduling and power allocation scheme for 6G terahertz mesh network is proposed in [135]. A mixed integer programming problem is formulated in the paper. A sub-optimal Greedy Shrinking Algorithms (GSA) is proposed in the paper. The GSA reduces the computational complexity of the optimisation problem.

I. 6G ENABLERS: IRS AND HOLOGRAPHIC MIMO

Intelligent Reflecting Surface (IRS) is considered as one of the key enablers for emerging 6G systems. It offers a variety of advantages including enhanced signal reception, increased secrecy rate and so on [136]. Holographic MIMO (HMIMO) surface is on the other hand, a cost efficient wireless planar structure which can transform or shape electromagnetic waves according to the requirement. They are also sometimes referred to as Large Intelligence Surface (LIS) [137]. Both IRS and HMIMO are expected to play a crucial role in enabling latency constraint services such as edge computing. In edge computing, Age of Information (AoI) is a critically important parameter that significantly affects the performance. AoI is highly correlated with latency. Various approaches are being considered for AoI minimisation such as using aerial IRS for improving signal quality at the destination [138]. The work applied the Successive Convex Approximation (SCA) algorithm for convergence and to minimise computational complexity which offers reduced latency. Another challenge in 6G systems is to ensure energy efficiency. In scheduling and routing, energy plays a crucial role. For example, depending on the signal quality of the receiver, scheduling priority changes. Thus IRS can be used to control the signal quality at the receiver and hence scheduling priority of the emerging 6G use cases such as air to ground mesh networks [139].

VIII. CONCLUSION

The emerging 6G network and systems are considered as heterogeneous services oriented technologies consisting of URLLC, eMBB and mMTC services. The requirements for each of the services are different. Hence designing the system that integrates the services into a single system is challenging especially the packet scheduling algorithm where the latency and reliability requirements of different services are different. In this paper we provide an extensive survey on recent advances and future outlook of packet scheduling algorithms in 5G and beyond systems. First, we have provided an overview of the scheduling algorithms with some important characteristics of the algorithms including throughput, link utilisation, delay bound consideration, fairness and complexity of the algorithms. Secondly, we have provided an overview of the metric based scheduling algorithms with a table that includes all the metrics of the algorithms. Thirdly, state-of-the-art descriptions of centralised and joint scheduling algorithms are presented. Finally, we provide research direction and future challenges of packet scheduling algorithms in 6G systems. The overviews and future directions presented in this survey provides an in-depth knowledge of the scheduling algorithms. From the discussion it is clear that significant research effort is required to design a fully functioning URLLC scheduling system for emerging 6G systems.

REFERENCES

[1] *IMT Vision–Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond*, document 2083, Recommendation ITU, 2015.

[2] F. Tariq, M. R. A. Khandaker, K.-K. Wong, M. A. Imran, M. Bennis, and M. Debbah, "A speculative study on 6G," *IEEE Wireless Commun.*, vol. 27, no. 4, pp. 118–125, Aug. 2020.

[3] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. Soong, and J. C. Zhang, "What will 5G be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

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

[5] F. Capozzi, G. Piro, L. A. Grieco, G. Boggia, and P. Camarda, "Downlink packet scheduling in LTE cellular networks: Key design issues and a survey," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 2, pp. 678–700, 2nd Quart., 2013.

[6] P. Popovski, "Ultra-reliable communication in 5G wireless systems," in *Proc. 1st Int. Conf. 5G Ubiquitous Connectivity*, 2014, pp. 146–151.

[7] S. Parkvall, E. Dahlman, A. Furuskar, and M. Frenne, "NR: The new 5G radio access technology," *IEEE Commun. Standards Mag.*, vol. 1, no. 4, pp. 24–30, Dec. 2017.

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

[9] M. M. Nasralla, N. Khan, and M. G. Martini, "Content-aware downlink scheduling for LTE wireless systems: A survey and performance comparison of key approaches," *Comput. Commun.*, vol. 130, pp. 78–100, Oct. 2018.

[10] R. Kwan and C. Leung, "A survey of scheduling and interference mitigation in LTE," *J. Electr. Comput. Eng.*, vol. 2010, pp. 1–10, Jan. 2010.

[11] M. A. Mehaseb, Y. Gadallah, A. Elhamy, and H. El-Hennawy, "Classification of LTE uplink scheduling techniques: An M2M perspective," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1310–1335, 2nd Quart., 2016.

[12] N. Abu-Ali, A. M. Taha, M. Salah, and H. Hassanein, "Uplink scheduling in LTE and LTE-advanced: Tutorial, survey and evaluation framework," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 3, pp. 1239–1265, 3rd Quart., 2014.

[13] T. A. Sheikh, J. Bora, and A. Hussain, "A survey of antenna and user scheduling techniques for massive MIMO-5G wireless system," in *Proc. Int. Conf. Current Trends Comput., Electr., Electron. Commun. (CTCEEC)*, Sep. 2017, pp. 578–583.

[14] A. Shahraki, M. Abbasi, M. J. Piran, and A. Taherkordi, "A comprehensive survey on 6G networks: Applications, core services, enabling technologies, and future challenges," 2021, *arXiv:2101.12475*.

[15] H. Tataria, M. Shafi, A. F. Molisch, M. Dohler, H. Sjöland, and F. Tufvesson, "6G wireless systems: Vision, requirements, challenges, insights, and opportunities," *Proc. IEEE*, vol. 109, no. 7, pp. 1166–1199, Jul. 2021.

[16] X. You et al., "Towards 6G wireless communication networks: Vision, enabling technologies, and new paradigm shifts," *Sci. China Inf. Sci.*, vol. 64, no. 1, pp. 1–74, Nov. 2020.

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

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

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

[20] S. R. Pokhrel, J. Ding, J. Park, O.-S. Park, and J. Choi, "Towards enabling critical mMTC: A review of URLLC within mMTC," *IEEE Access*, vol. 8, pp. 131796–131813, 2020.

[21] C. Sun, C. She, and C. Yang, "Retransmission policy with frequency hopping for ultra-reliable and low-latency communications," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2018, pp. 1–6.

[22] R. Abreu, P. Mogensen, and K. I. Pedersen, "Pre-scheduled resources for retransmissions in ultra-reliable and low latency communications," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Mar. 2017, pp. 1–5.

[23] H. Fattah and C. Leung, "An overview of scheduling algorithms in wireless multimedia networks," *IEEE Wireless Commun.*, vol. 9, no. 5, pp. 76–83, Oct. 2002.

[24] *Study on Latency Reduction Techniques for LTE*, 3GPP, document TR 36.881, 2015.

[25] *Study on Scenarios and Requirements for Next Generation Access Technologies*, 3GPP, document TR 38.913, 2018.

[26] B. Soret, P. Mogensen, K. I. Pedersen, and M. C. Aguayo-Torres, "Fundamental tradeoffs among reliability, latency and throughput in cellular networks," in *Proc. IEEE Globecom Workshops*, Dec. 2014, pp. 1391–1396.

[27] *Verticals URLLC Use Cases and Requirements*, NGMN Alliance, Frankfurt am Main, Germany, 2019.

[28] J.-P. Stauffert, F. Niebling, and M. E. Latoschik, "Latency and cybersickness: Impact, causes, and measures. A review," *Frontiers Virtual Reality*, vol. 1, Nov. 2020, Art. no. 582204.

[29] Q. Zhang, J. Liu, and G. Zhao, "Towards 5G enabled tactile robotic telesurgery," 2018, *arXiv:1803.03586*.

[30] R. Gupta, S. Tanwar, S. Tyagi, and N. Kumar, "Tactile-internet-based telesurgery system for healthcare 4.0: An architecture, research challenges, and future directions," *IEEE Netw.*, vol. 33, no. 6, pp. 22–29, Nov. 2019.

[31] T. M. Ho, T. D. Tran, T. T. Nguyen, S. M. A. Kazmi, L. B. Le, C. S. Hong, and L. Hanzo, "Next-generation wireless solutions for the smart factory, smart vehicles, the smart grid and smart cities," 2019, *arXiv:1907.10102*.

[32] B. S. Tsybakov, "File transmission over wireless fast fading downlink," *IEEE Trans. Inf. Theory*, vol. 48, no. 8, pp. 2323–2337, Aug. 2002.

[33] P. Kela, J. Puttonen, N. Kolehmainen, T. Ristaniemi, T. Henttonen, and M. Moisio, "Dynamic packet scheduling performance in UTRA long term evolution downlink," in *Proc. 3rd Int. Symp. Wireless Pervasive Comput.*, May 2008, pp. 308–313.

[34] R. Basukala, H. M. Ramli, and K. Sandrasegaran, "Performance analysis of EXP/PF and M-LWDF in downlink 3GPP LTE system," in *Proc. 1st Asian Himalayas Int. Conf. Internet*, Nov. 2009, pp. 1–5.

[35] S. Riahi and A. Riahi, "Optimal performance of the opportunistic scheduling in new generation mobile systems," in *Proc. 2nd Int. Conf. Smart Digit. Environ.*, Oct. 2018, pp. 33–37.

[36] M. A. Ibraheem, N. ElShennawy, and A. M. Sarhan, "A proposed modified proportional fairness scheduling (MPF-BCQI) algorithm with best CQI consideration for LTE—A networks," in *Proc. 13th Int. Conf. Comput. Eng. Syst. (ICCES)*, Dec. 2018, pp. 360–368.

[37] M.-R. Hojeij, C. A. Nour, J. Farah, and C. Douillard, "Weighted proportional fair scheduling for downlink nonorthogonal multiple access," *Wireless Commun. Mobile Comput.*, vol. 2018, pp. 1–12, Jan. 2018.

[38] E. Aryafar, A. Keshavarz-Haddad, and C. Joe-Wong, "Proportional fair RAT aggregation in HetNets," in *Proc. 31st Int. Teletraffic Congr.*, Aug. 2019, pp. 84–92.

[39] M. Ismail, A. Isa, M. Johal, M. Ahmad, M. Zin, M. I. M. Isa, H. Nornikman, and M. Mahyuddin, "Design and analysis of modified-proportional fair scheduler for LTE femtocell networks," *J. Telecommun., Electron. Comput. Eng.*, vol. 9, nos. 2–5, pp. 7–11, 2017.

[40] I. M. Delgado-Luque, M. C. Aguayo-Torres, G. Gómez, F. J. Martín-Vega, and J. T. Entrambasaguas, "SNR-versus rate-based proportional fair scheduling in Rayleigh fading channels," *Wireless Pers. Commun.*, vol. 106, no. 4, pp. 2099–2111, Jun. 2019.

[41] J.-H. Rhee, J. M. Holtzman, and D.-K. Kim, "Scheduling of real/non-real time services: Adaptive EXP/PF algorithm," in *Proc. 57th IEEE Semiannual Veh. Technol. Conf.*, Apr. 2003, pp. 462–466.

[42] B. Sadiq, R. Madan, and A. Sampath, "Downlink scheduling for multiclass traffic in LTE," *EURASIP J. Wireless Commun. Netw.*, vol. 2009, no. 1, pp. 1–18. Mar. 2009.

[43] K. Ramanan and A. L. Stolyar, "Largest weighted delay first scheduling: Large deviations and optimality," *Ann. Appl. Probab.*, vol. 11, no. 1, pp. 1–48, Feb. 2001.

[44] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, and R. Vijayakumar, "Providing quality of service over a shared wireless link," *IEEE Commun. Mag.*, vol. 39, no. 2, pp. 150–154, Feb. 2001.

[45] H. A. M. Ramli, R. Basukala, K. Sandrasegaran, and R. Patachaianand, "Performance of well known packet scheduling algorithms in the downlink 3GPP LTE system," in *Proc. IEEE 9th Malaysia Int. Conf. Commun. (MICC)*, Dec. 2009, pp. 815–820.

[46] M. I. Husain, M. E. Haque, and F. Tariq, "An efficient packet scheduling algorithm for URLLC systems," in *Proc. Int. Conf. UK-China Emerg. Technol. (UCET)*, Aug. 2020, pp. 1–4.

[47] A. Destounis, G. S. Paschos, J. Arnau, and M. Kountouris, "Scheduling URLLC users with reliable latency guarantees," in *Proc. 16th Int. Symp. Modeling Optim. Mobile, Ad Hoc, Wireless Netw. (WiOpt)*, May 2018, pp. 1–8.

[48] N. B. Khalifa, M. Assaad, and M. Debbah, "Risk-sensitive reinforcement learning for URLLC traffic in wireless networks," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2019, pp. 1–7.

[49] W. J. Ryu and S. Y. Shin, "Performance evaluation of a power allocation algorithm based on dynamic blocklength estimation for URLLC in the multicarrier downlink NOMA systems," *TURKISH J. Electr. Eng. Comput. Sci.*, vol. 29, no. 1, pp. 310–320, Jan. 2021.

[50] N. B. Khalifa, V. Angilella, M. Assaad, and M. Debbah, "Low-complexity channel allocation scheme for URLLC traffic," *IEEE Trans. Commun.*, vol. 69, no. 1, pp. 194–206, Jan. 2021.

[51] S. He, Z. An, J. Zhu, J. Zhang, Y. Huang, and Y. Zhang, "Beamforming design for multiuser URLLC with finite blocklength transmission," *IEEE Trans. Wireless Commun.*, vol. 20, no. 12, pp. 8096–8109, Dec. 2021.

[52] P. Caballero, A. Banchs, G. de Veciana, X. Costa-Pérez, and A. Azcorra, "Network slicing for guaranteed rate services: Admission control and resource allocation games," *IEEE Trans. Wireless Commun.*, vol. 17, no. 10, pp. 6419–6432, Oct. 2018.

[53] S. Li, N. Zhang, S. Lin, L. Kong, A. Katangur, M. K. Khan, M. Ni, and G. Zhu, "Joint admission control and resource allocation in edge computing for Internet of Things," *IEEE Netw.*, vol. 32, no. 1, pp. 72–79, Feb. 2018.

[54] A. Destounis and G. S. Paschos, "Complexity of URLLC scheduling and efficient approximation schemes," in *Proc. Int. Symp. Modeling Optim. Mobile, Ad Hoc, Wireless Netw. (WiOPT)*, Jun. 2019, pp. 1–8.

[55] Z. Meng and H. Zhang, "Multi-cell, multi-channel scheduling with probabilistic per-packet real-time guarantee," 2021, *arXiv:2101.01768*.

[56] Y. Chen, H. Zhang, N. Fisher, L. Y. Wang, and G. Yin, "Probabilistic per-packet real-time guarantees for wireless networked sensing and control," *IEEE Trans. Ind. Informat.*, vol. 14, no. 5, pp. 2133–2145, May 2018.

[57] C. Wu, M. Sha, D. Gunatilaka, A. Saifullah, C. Lu, and Y. Chen, "Analysis of EDF scheduling for wireless sensor-actuator networks," in *Proc. IEEE 22nd Int. Symp. Quality Service (IWQoS)*, May 2014, pp. 31–40.

[58] D. Han and W. Chen, "QoS differential scheduling of URLLC under FIFO service discipline: A cross-layer approach," *IEEE Wireless Commun. Lett.*, vol. 9, no. 9, pp. 1370–1373, Sep. 2020.

[59] D. Han, W. Chen, and Y. Fang, "Joint channel and queue aware scheduling for latency sensitive mobile edge computing with power constraints," *IEEE Trans. Wireless Commun.*, vol. 19, no. 6, pp. 3938–3951, Jun. 2020.

[60] X. Zhao, W. Chen, J. Lee, and N. B. Shroff, "Delay-optimal and energy-efficient communications with Markovian arrivals," *IEEE Trans. Commun.*, vol. 68, no. 3, pp. 1508–1523, Mar. 2019.

[61] A. Nassar and Y. Yilmaz, "Reinforcement learning for adaptive resource allocation in fog RAN for IoT with heterogeneous latency requirements," *IEEE Access*, vol. 7, pp. 128014–128025, 2019.

[62] J. Cheng, C. Shen, and S. Xia, "Robust URLLC packet scheduling of OFDM systems," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, May 2020, pp. 1–6.

[63] L. Zhu, L. Feng, Z. Yang, W. Li, and Q. Ou, "Priority-based uRLLC uplink resource scheduling for smart grid neighborhood area network," in *Proc. IEEE Int. Conf. Energy Internet (ICEI)*, May 2019, pp. 510–515.

[64] S. Wu, Z. Wang, Z. Li, Z. WeiJun, W. Shao, B. Ma, S. Yao, and Y. Wang, "Uplink resource allocation based on short block-length regime in heterogeneous cellular networks for smart grid," in *Proc. Int. Conf. Innov. Mobile Internet Services Ubiquitous Comput.* Cham, Switzerland: Springer, 2020, pp. 213–224.

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

[66] A. Elgabli, H. Khan, M. Krouka, and M. Bennis, "Reinforcement learning based scheduling algorithm for optimizing age of information in ultra reliable low latency networks," in *Proc. IEEE Symp. Comput. Commun. (ISCC)*, Jun. 2019, pp. 1–6.

[67] A. Yazar and H. Arslan, "Reliability enhancement in multi-numerology-based 5G new radio using INI-aware scheduling," *EURASIP J. Wireless Commun. Netw.*, vol. 2019, no. 1, pp. 1–14, Dec. 2019.

[68] A. Karimi, K. I. Pedersen, N. H. Mahmood, J. Steiner, and P. Mogensen, "5G centralized multi-cell scheduling for URLLC: Algorithms and system-level performance," *IEEE Access*, vol. 6, pp. 72253–72262, 2018.

[69] A. Karimi, K. I. Pedersen, and P. Mogensen, "Low-complexity centralized multi-cell radio resource allocation for 5G URLLC," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, May 2020, pp. 1–6.

[70] A. Karimi, K. I. Pedersen, N. H. Mahmood, J. Steiner, and P. Mogensen, "Centralized joint cell selection and scheduling for improved URLLC performance," in *Proc. IEEE 29th Annu. Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC)*, Sep. 2018, pp. 1–6.

[71] J. J. Escudero-Garzás, C. Bousoño-Calzón, and A. García, "On the feasibility of 5G slice resource allocation with spectral efficiency: A probabilistic characterization," *IEEE Access*, vol. 7, pp. 151948–151961, 2019.

[72] M. Pagin, T. Zugno, M. Polese, and M. Zorzi, "Resource management for 5G NR integrated access and backhaul: A semi-centralized approach," *IEEE Trans. Wireless Commun.*, vol. 21, no. 2, pp. 753–767, Feb. 2022.

[73] T. Sakai, Y. Yuda, and K. Higuchi, "Inter-base station cooperative scheduling method among multiple service channels to maximize integrated system throughput," in *Proc. 21st Int. Symp. Wireless Pers. Multimedia Commun. (WPMC)*, Nov. 2018, pp. 273–277.

[74] D. Öhmann, "High reliability in wireless networks through multi-connectivity," Ph.D. thesis, Technische Univ. Dresden, 2017.

[75] A. Wolf, P. Schulz, M. Dörpinghaus, J. C. S. S. Filho, and G. Fettweis, "How reliable and capable is multi-connectivity?" *IEEE Trans. Commun.*, vol. 67, no. 2, pp. 1506–1520, Feb. 2019.

[76] R. P. Antonioli, J. Pettersson, and T. F. Maciel, ''Split responsibility scheduler for multi-connectivity in 5G cellular networks,'' *IEEE Netw.*, vol. 34, no. 6, pp. 212–219, Nov. 2020.

[77] Y. Wang, K. I. Pedersen, T. B. Sórensen, and P. E. Mogensen, ''Carrier load balancing and packet scheduling for multi-carrier systems,'' *IEEE Trans. Wireless Commun.*, vol. 9, no. 5, pp. 1780–1789, May 2010.

[78] V. F. Monteiro, D. A. Sousa, T. F. Maciel, F. R. P. Cavalcanti, C. F. M. E. Silva, and E. B. Rodrigues, ''Distributed RRM for 5G multi-RAT multiconnectivity networks,'' *IEEE Syst. J.*, vol. 13, no. 1, pp. 192–203, Mar. 2019.

[79] M. Peng, C. Wang, V. Lau, and H. V. Poor, ''Fronthaul-constrained cloud radio access networks: Insights and challenges,'' *IEEE Wireless Commun.*, vol. 22, no. 2, pp. 152–160, Apr. 2015.

[80] B. Rouzbehani, V. Marbukh, K. Sayrafian, and L. M. Correia, ''Towards cross-layer optimization of virtualized radio access networks,'' in *Proc. Eur. Conf. Netw. Commun. (EuCNC)*, Jun. 2019, pp. 448–452.

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

[82] S. Skaperas, N. Ferdosian, A. Chorti, and L. Mamatas, ''Scheduling optimization of heterogeneous services by resolving conflicts,'' 2021, *arXiv:2103.01897*.

[83] L. You, Q. Liao, N. Pappas, and D. Yuan, ''Resource optimization with flexible numerology and frame structure for heterogeneous services,'' *IEEE Commun. Lett.*, vol. 22, no. 12, pp. 2579–2582, Dec. 2018.

[84] Q. He, G. Dan, and G. P. Koudouridis, ''Semi-persistent scheduling for 5G downlink based on short-term traffic prediction,'' in *Proc. IEEE Global Commun. Conf.*, Dec. 2020, pp. 1–6.

[85] C. Li, J. Li, and W. Chen, ''Adaptive ultra-reliable low-latency communications (URLLC) semi-persistent scheduling,'' U.S. Patent 15/388 512, Jan. 28, 2018.

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

[87] R. Kassab, O. Simeone, and P. Popovski, ''Coexistence of URLLC and eMBB services in the C-RAN uplink: An information-theoretic study,'' in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2018, pp. 1–6.

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

[89] A. A. Esswie and K. I. Pedersen, ''Multi-user preemptive scheduling for critical low latency communications in 5G networks,'' in *Proc. IEEE Symp. Comput. Commun. (ISCC)*, Jun. 2018, pp. 136–141.

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

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

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

[93] F. B. Tesema, A. Awada, I. Viering, M. Simsek, and G. P. Fettweis, ''Mobility modeling and performance evaluation of multi-connectivity in 5G intra-frequency networks,'' in *Proc. IEEE Globecom Workshops*, Dec. 2015, pp. 1–6.

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

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

[96] M. Morcos, M. Mhedhbi, A. Galindo-Serrano, and S. Eddine Elayoubi, ''Optimal resource preemption for aperiodic URLLC traffic in 5G networks,'' in *Proc. IEEE 31st Annu. Int. Symp. Pers., Indoor Mobile Radio Commun.*, Aug. 2020, pp. 1–6.

[97] W. Zhang, M. Derakhshani, and S. Lambotharan, ''Stochastic optimization of URLLC-eMBB joint scheduling with queuing mechanism,'' *IEEE Wireless Commun. Lett.*, vol. 10, no. 4, pp. 844–848, Apr. 2021.

[98] N. Ksairi and M. Kountouris, ''Timely scheduling of URLLC packets using precoder compatibility estimates,'' in *Proc. IEEE 30th Annu. Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC)*, Sep. 2019, pp. 1–6.

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

[100] M. Almekhlafi, M. A. Arfaoui, C. Assi, and A. Ghrayeb, ''Joint resource and power allocation for URLLC-eMBB traffics multiplexing in 6G wireless networks,'' in *Proc. IEEE Int. Conf. Commun.*, Jun. 2021, pp. 1–6.

[101] O. A. Nwogu, G. Diaz, and M. Abdennebi, ''Differential traffic QoS scheduling for 5G/6G fronthaul networks,'' in *Proc. 31st Int. Telecommun. Netw. Appl. Conf. (ITNAC)*, Nov. 2021, pp. 113–120.

[102] L. Feng, Y. Zi, W. Li, F. Zhou, P. Yu, and M. Kadoch, ''Dynamic resource allocation with RAN slicing and scheduling for uRLLC and eMBB hybrid services,'' *IEEE Access*, vol. 8, pp. 34538–34551, 2020.

[103] S. O. Oladejo and O. E. Falowo, ''Latency-aware dynamic resource allocation scheme for multi-tier 5G network: A network slicing-multitenancy scenario,'' *IEEE Access*, vol. 8, pp. 74834–74852, 2020.

[104] S. E. Elayoubi, P. Brown, M. Deghel, and A. Galindo-Serrano, ''Radio resource allocation and retransmission schemes for URLLC over 5G networks,'' *IEEE J. Sel. Areas Commun.*, vol. 37, no. 4, pp. 896–904, Apr. 2019.

[105] Y. Wu, D. Wu, L. Ao, L. Yang, and Q. Fu, ''Contention-based radio resource management for URLLC-oriented D2D communications,'' *IEEE Trans. Veh. Technol.*, vol. 69, no. 9, pp. 9960–9971, Sep. 2020.

[106] W. Ning, Y. Wang, M. Liu, Y. Chen, and X. Wang, ''Mission-critical resource allocation with puncturing in industrial wireless networks under mixed services,'' *IEEE Access*, vol. 9, pp. 21870–21880, 2021.

[107] Y. Huang, S. Li, C. Li, Y. T. Hou, and W. Lou, ''A deep-reinforcement-learning-based approach to dynamic eMBB/URLLC multiplexing in 5G NR,'' *IEEE Internet Things J.*, vol. 7, no. 7, pp. 6439–6456, Jul. 2020.

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

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

[110] H. Khan, M. M. Butt, S. Samarakoon, P. Sehier, and M. Bennis, ''Deep learning assisted CSI estimation for joint URLLC and eMBB resource allocation,'' in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, Jun. 2020, pp. 1–6.

[111] F. Saggese, L. Pasqualini, M. Moretti, and A. Abrardo, ''Deep reinforcement learning for URLLC data management on top of scheduled eMBB traffic,'' 2021, *arXiv:2103.01801*.

[112] Y. Li, C. Hu, J. Wang, and M. Xu, ''Optimization of URLLC and eMBB multiplexing via deep reinforcement learning,'' in *Proc. IEEE/CIC Int. Conf. Commun. Workshops China*, Aug. 2019, pp. 245–250.

[113] M. Alsenwi, N. H. Tran, M. Bennis, S. R. Pandey, A. K. Bairagi, and C. S. Hong, ''Intelligent resource slicing for eMBB and URLLC coexistence in 5G and beyond: A deep reinforcement learning based approach,'' *IEEE Trans. Wireless Commun.*, vol. 20, no. 7, pp. 4585–4600, Jul. 2021.

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

[115] E. Khorov, A. Krasilov, and A. Malyshev, ''Radio resource and traffic management for ultra-reliable low latency communications,'' in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2018, pp. 1–6.

[116] J. He, K. Yang, and H.-H. Chen, ''6G cellular networks and connected autonomous vehicles,'' *IEEE Netw.*, vol. 35, no. 4, pp. 255–261, Jul./Aug. 2021.

[117] J. Gao, M. R. A. Khandaker, F. Tariq, K.-K. Wong, and R. T. Khan, ''Deep neural network based resource allocation for V2X communications,'' in *Proc. IEEE 90th Veh. Technol. Conf. (VTC-Fall)*, Sep. 2019, pp. 1–5.

[118] D. C. Nguyen, M. Ding, P. N. Pathirana, A. Seneviratne, J. Li, D. Niyato, O. Dobre, and H. V. Poor, ''6G Internet of Things: A comprehensive survey,'' *IEEE Internet Things J.*, vol. 9, no. 1, pp. 359–383, Jan. 2022.

[119] T. Zhang, ''Resource allocation for next generation wireless networks,'' Ph.D. thesis, Dept. Elect. Comput. Eng., Auburn Univ., Auburn, AL, USA, 2022.

[120] H. Ren, C. Pan, Y. Deng, M. Elkashlan, and A. Nallanathan, ''Resource allocation for secure URLLC in mission-critical IoT scenarios,'' *IEEE Trans. Commun.*, vol. 68, no. 9, pp. 5793–5807, Sep. 2020.

[121] S. A. Ashraf, F. Lindqvist, R. Baldemair, and B. Lindoff, ''Control channel design trade-offs for ultra-reliable and low-latency communication system,'' in *Proc. IEEE Globecom Workshops*, Dec. 2015, pp. 1–6.

[122] W. Kim and B. Shim, "Ultra-mini slot transmission for 5G+ and 6G URLLC network," in *Proc. IEEE 92nd Veh. Technol. Conf.*, Nov. 2020, pp. 1–5.

[123] J. M. C. Knot, J. C. M. van den Ende, and P. J. Vergragt, "Flexibility strategies for sustainable technology development," *Technovation*, vol. 21, no. 6, pp. 335–343, Jun. 2001.

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

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

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

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

[128] O. N. C. Yilmaz, O. Teyeb, and A. Orsino, "Overview of LTE-NR dual connectivity," *IEEE Commun. Mag.*, vol. 57, no. 6, pp. 138–144, Dec. 2019.

[129] S. Singh and I. Chana, "A survey on resource scheduling in cloud computing: Issues and challenges," *J. Grid Comput.*, vol. 14, no. 2, pp. 217–264, Jun. 2016.

[130] G. Pocovi, H. Shariatmadari, G. Berardinelli, K. Pedersen, J. Steiner, and Z. Li, "Achieving ultra-reliable low-latency communications: Challenges and envisioned system enhancements," *IEEE Netw.*, vol. 32, no. 2, pp. 8–15, Mar./Apr. 2018.

[131] M. Hao, D. Ye, S. Wang, B. Tan, and R. Yu, "URLLC resource slicing and scheduling for trustworthy 6G vehicular services: A federated reinforcement learning approach," *Phys. Commun.*, vol. 49, Dec. 2021, Art. no. 101470.

[132] T.-H. Li, M. R. A. Khandaker, F. Tariq, K.-K. Wong, and R. T. Khan, "Learning the wireless V2I channels using deep neural networks," in *Proc. IEEE 90th Veh. Technol. Conf. (VTC-Fall)*, Sep. 2019, pp. 1–5.

[133] G.-X. Shen, M. R. A. Khandaker, and F. Tariq, "Learning the wireless channel: A deep neural network approach," in *Proc. Int. Conf. UK-China Emerg. Technol. (UCET)*, Aug. 2020, pp. 1–6.

[134] K. B. Letaief, Y. Shi, J. Lu, and J. Lu, "Edge artificial intelligence for 6G: Vision, enabling technologies, and applications," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 5–36, Jan. 2022.

[135] M. Yu, A. Tang, X. Wang, and C. Han, "Joint scheduling and power allocation for 6G terahertz mesh networks," in *Proc. Int. Conf. Comput., Netw. Commun. (ICNC)*, Feb. 2020, pp. 631–635.

[136] Y. Song, M. R. A. Khandaker, F. Tariq, K.-K. Wong, and A. Toding, "Truly intelligent reflecting surface-aided secure communication using deep learning," in *Proc. IEEE 93rd Veh. Technol. Conf. (VTC-Spring)*, Apr. 2021, pp. 1–6.

[137] C. Huang, S. Hu, G. C. Alexandropoulos, A. Zappone, C. Yuen, R. Zhang, M. D. Renzo, and M. Debbah, "Holographic MIMO surfaces for 6G wireless networks: Opportunities, challenges, and trends," *IEEE Wireless Commun.*, vol. 27, no. 5, pp. 118–125, Oct. 2020.

[138] W. Lyu, Y. Xiu, S. Yang, P. L. Yeoh, and Y. Li, "Weighted sum age of information minimization in wireless networks with aerial IRS," 2022, *arXiv:2203.04525*.

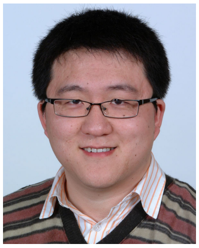
[139] Y. Liu, H. Mao, L. Zhu, Z. Xiao, Z. Han, and X.-G. Xia, "Routing and resource scheduling for air-ground integrated mesh networks," *IEEE Trans. Wireless Commun.*, early access, Nov. 28, 2022, doi: 10.1109/TWC.2022.3223152.

MD. EMDADUL HAQUE (Member, IEEE) received the B.Sc. degree in computer science and engineering from Khulna University, Bangladesh, the M.Sc. degree in computer science and engineering from the Bangladesh University of Engineering and Technology, and the Ph.D. degree from the Tokyo Institute of Technology, Japan. He is currently a Professor of information and communication engineering with the University of Rajshahi. His research interests include 5G wireless networks, wireless sensor networks, and machine learning.

FAISAL TARIQ (Senior Member, IEEE) is currently a Senior Lecturer with the James Watt School of Engineering, University of Glasgow, U.K. Before this, he was a Lecturer with the Queen Mary University of London. His research interests include 56/6G wireless communications, physical layer security, and machine-learning applications. He received the Best Paper Award from IEEE WCMC 2013. He is serving as an Associate Editor for IEEE Wireless Communications Letters and the Editor for Journal of Networks and Computer Applications (Elsevier).

MUHAMMAD R. A. KHANDAKER (Senior Member, IEEE) is currently an Assistant Professor with the School of Engineering and Physical Sciences, Heriot-Watt University. Before joining Heriot-Watt University, he was a Postdoctoral Research Fellow with University College London, U.K., from July 2013 to June 2018. He is an Associate Editor of the IEEE Wireless Communications Letters, IEEE Communications Letters, IEEE Access, and *EURASIP Journal on Wireless Communications and Networking*.

KAI-KIT WONG (Fellow, IEEE) received the B.Eng., M.Phil., and Ph.D. degrees in electrical and electronic engineering from The Hong Kong University of Science and Technology, Hong Kong, in 1996, 1998, and 2001, respectively. After graduation, he took up academic and research positions with The University of Hong Kong; Lucent Technologies, Bell-Labs, Holmdel; the Smart Antennas Research Group, Stanford University; and the University of Hull, U.K. He is currently the Chair of wireless communications with the Department of Electronic and Electrical Engineering, University College London, U.K. His current research interest includes 5G and beyond mobile communications. He is a fellow of IET. He was a co-recipient of the 2013 IEEE Signal Processing Letters Best Paper Award; the 2000 IEEE VTS Japan Chapter Award from the IEEE Vehicular Technology Conference, Japan, in 2000; and a few other international best paper awards. He is on the editorial board of several international journals. He has been the Editor-in-Chief of IEEE Wireless Communications Letters, since 2020.



YANGYANG ZHANG received the B.S. and M.S. degrees in electronics and information engineering from Northeastern University, Shenyang, China, in 2002 and 2004, respectively, and the Ph.D. degree in electrical engineering from the University of Oxford, in 2008. From 2008 to 2010, he was a Postdoctoral Research Fellow with University College London. He is currently with the Shenzhen Key Laboratory of Artificial Microstructure Design, Guangdong Key Laboratory of Meta-RF Microwave Radio Frequency, and the Kuang-Chi Institute of Advanced Technology, Shenzhen, China. He is also the Executive Vice President with the Kuang-Chi Institute of Advanced Technology. His research interests include metamaterial-based future wireless communication systems, such as MIMO communication systems, metamaterial-based RF devices, and metamaterial-based spatial modulation technology. He received more than 20 honors from various national and international competitions.