Real-Time Flow Scheduling in Industrial 5G New Radio

Tianyu Zhang*, Jiachen Wang*, Xiaobo Sharon Hu†, Song Han*
*Dept. of Computer Science and Engineering, University of Connecticut
*Email: {tianyu.zhang, jc.wang, song.han}@uconn.edu
†Dept. of Computer Science and Engineering, University of Notre Dame
†Email: shu@nd.edu

***Abstract*—Among the many industrial wireless solution candidates, 5G New Radio (NR) has drawn significant attention in recent years due to its capabilities to support ultra-high-speed communication, ultra-low latency, and massive connectivity. Despite its great potential, 5G NR also brings significant complexity in scheduling industrial data flows to meet their hard real-time requirements. In this paper, we first leverage a real-world 5G RAN testbed to benchmark the downlink throughput and explore the impact of modulation and coding scheme (MCS) selection on the network performance. We then formulate a real-time flow scheduling problem in industrial 5G NR, which features per-flow real-time schedulability guarantees through time-frequency-space resource allocation. We propose a novel two-phase scheduling framework, named 5G-TPS, to construct the schedule that meets the deadlines of all the flows. To adapt to dynamic channel conditions, 5G-TPS enables online schedule adjustment for affected flows to meet their timing requirements. To evaluate the performance of 5G-TPS, we present a case study of a motion control panel use case and perform extensive experiments. The results show that 5G-TPS can achieve schedulability ratios comparable to the Satisfiability Modulo Theory (SMT)-based exact solution and outperform many other state-of-the-art scheduling approaches, including the built-in 5G NR schedulers.**

I. Introduction

Industrial Internet-of-Things (IIoT) is expected to significantly improve the efficiency and performance of industrial networks across a wide range of industrial applications. Many of these industrial applications (e.g., use cases specified by 3GPP [1] including mobile operation panels and remote surgery) are mission- and safety-critical, with stringent timing and reliability requirements on the communication fabric to exchange information among various devices [2].

IIoT tends to use wireless networks for communication since they enable more flexible network configurations and reduce cabling costs compared to their wired counterparts (e.g., industrial Ethernet [3] and Time-Sensitive Networking [4]). However, existing industrial wireless solutions (e.g., ISA100.11, WirelessHART, and 6TiSCH [5]–[8]) are mainly used in the context of low-power and low-speed wireless sensor and actuator networks. To support high-speed real-time wireless communication, IEEE 802.11-based protocols (e.g., Wi-Fi 6 [9]) have received growing attention in industrial applications due to their low deployment cost. However, 802.11-based protocols operate in unlicensed spectrum and may suffer severe and unexpected interference from other co-existing networks.

The industrial connectivity landscape is changing with the emergence of 5G New Radio (NR) cellular networks [10]. The deployment of 5G NR in industrial applications, also termed private 5G networks in 3GPP, has attracted significant interest due to its capabilities of providing ultra-high-speed communication (multi-Gbps peak rates), wide coverage, ultra-low latency, and massive connectivity. Furthermore, the private 5G deployment options also provide complete control to configure every aspect of the network (e.g., schedule, resource allocation) without involving mobile network operators [2].

To achieve ultra high-speed real-time communication, several enabling technologies are supported in industrial 5G NR. For example, orthogonal frequency division multiple access (OFDMA) is utilized in 5G NR for both uplink (UL) and downlink (DL) to achieve deterministic transmissions [11]. Compared to 4G LTE networks, 5G NR adopts fewer OFDM symbols per transmission time interval (TTI) and shortens the OFDM symbols via a wider subcarrier spacing to reduce latency. MU-MIMO (multi-user multiple-input-multiple-output) is another core technology for 5G NR to significantly increase network throughput [12] by allowing a base station gNB to harvest the spatial diversity and transmit signals to multiple user equipment (UEs) on the same frequency band simultaneously. In addition, 5G NR provides robust modulation and coding schemes (MCS) which determines the user's data rate on individual frequency bands according to the per-band channel quality indicator (CQI), namely subband CQI report.

Although 5G NR provides much flexibility and has tremendous potential, it brings high complexity due to the large design space of the flow scheduler at the gNB to meet the real-time requirements of industrial data flows. Specifically, the scheduler needs to i) allocate resource blocks (RBs) in the frequency domain to multiple users appropriately; ii) determine the number of data streams that can be transmitted by each user simultaneously in the spatial domain based on their transmission interference; and iii) choose the MCS index for each user ensuring that the selected MCS index is identical across all RBs allocated to this user [13]. Therefore, the real-time flow scheduling problem in 5G NR couples together RB allocation, data stream number determination, and MCS index selection in order to satisfy the timing requirements of industrial data flows, making the problem extremely challenging.

Time-frequency scheduling for real-time flows in traditional industrial wireless networks has been well studied [14]–[18]. In 5G networks, many recent works studied resource schedul-

2023 IEEE Real-Time Systems Symposium (RTSS) | 979-8-3503-2857-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/RTSS59052.2023.00039

2576-3172/23/$31.00 ©2023 IEEE
DOI 10.1109/RTSS59052.2023.00039

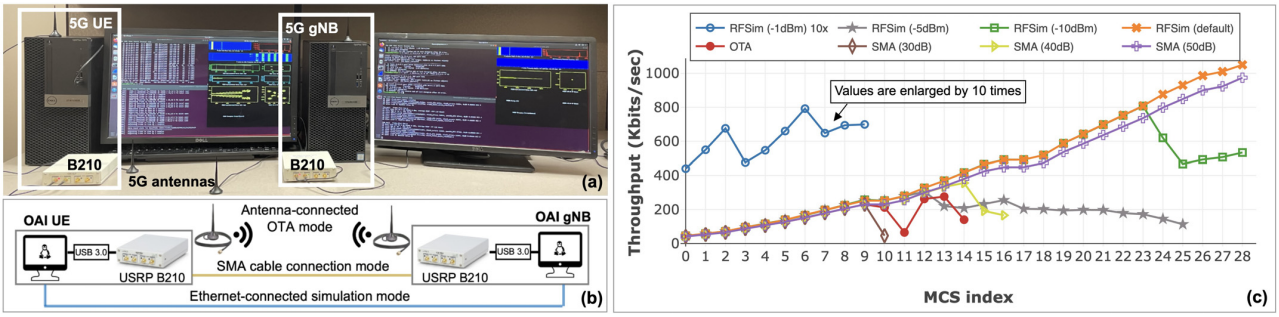

Fig. 1. Motivational experiments on a 5G RAN testbed. (a) Overview of the OAI-based 5G RAN testbed consisting of one gNB and one UE. (b) Architecture of the 5G RAN testbed with three connection modes. (c) Throughput results with varied MCS indices under various connectivity settings.

ing problems with objectives to optimize network throughput (e.g., [19]–[22]) or AoI (Age of Information, e.g., [23]–[25]). In light of improving the real-time performance of 5G networks, [26], [27] provide hard performance guarantees through formal response time analysis for 5G network slicing. However, these works adopt over-simplified resource models; and the proposed analyses only apply to fixed-priority scheduling, which leads to low schedulability performance as revealed in our experimental results. Some recent works [28]–[30] study 5G configured grant (CG) scheduling aiming at providing real-time guarantees for time-critical traffic. However, CG scheduling is only applied to 5G UL transmissions and suffers from extremely low flexibility. To the best of our knowledge, this paper is the first work that studies time-frequency-space DL scheduling in industrial 5G NR, which features per-flow real-time schedulability guarantee. Specifically, we make the following contributions.

- We leverage a real-world 5G RAN testbed to benchmark the DL throughput and explore the impact of MCS index selection on the network performance.
- We formulate the real-time flow scheduling problem in industrial 5G NR considering the featured 5G techniques, e.g., MU-MIMO and MCS selection based on subband CQI report.
- We introduce a two-phase scheduling framework, 5G-TPS, to construct a feasible schedule with deadline guarantees for all the flows. Upon any dynamic channel condition changes, 5G-TPS enables online schedule adjustment for affected flows to meet their timing requirements.
- We evaluate the performance of 5G-TPS through a case study and extensive experiments by comparing it with an SMT-based exact solution and many other state-of-the-art methods. The results demonstrate superior performance of 5G-TPS in terms of schedulability ratio, in both stable and dynamic channel conditions.

II. Motivational Experiments

In traditional industrial wireless networks, considerable research has been conducted on channel allocation in the frequency domain and flow scheduling in the time domain (e.g., [31]–[33]). However, the MCS index selection in 5G NR scheduling remains an area that lacks comprehensive understanding [34], [35], particularly regarding its impact on the network performance.

In 5G NR, MCS determines the number of bits per symbol that can be modulated and coded on the transmission channel between UE and the gNB. Higher MCS indices generally correspond to higher channel efficiency and higher data rates, but they may also be associated with higher packet error rates, as more aggressive modulation schemes are more susceptible to noise and interference. In order to understand the impact of the selected MCS index on the performance of 5G NR, we constructed a real-world 5G RAN testbed using OpenAirInterface (OAI) [36], a widely adopted open-source project which provides 3GPP-compliant implementations of gNB and UE, and benchmarked the downlink (DL) throughput of the OAI-based 5G RAN for different MCS index values and connectivity settings.

A. Testbed Setup

Our 5G RAN testbed, as shown in Fig. 1(a), comprises one gNB and one UE, each of which runs the OAI stack on a host machine (Intel i7-9700 processor @3.00GHz, 8 Cores, 64 GB RAM). Each host machine is connected to a USRP B210 device via USB 3.0, serving as the radio head unit (RHU).

1) Connectivity Settings: To thoroughly study the impact of the MCS index on network throughput, we conduct experiments in three different connection modes as shown in Fig. 1(b). In the *RFSim* mode, the OAI gNB transmits the I/Q samples to the UE over a radio channel simulator, namely RF simulator, via Ethernet without using the RF boards (i.e., USRP B210). In the *over-the-air (OTA)* mode, omnidirectional antennas are connected to the RF boards to transmit signals. In the *SMA cable* mode, the two USRP B210 devices are directly connected using Sub-Miniature version A (SMA) cables instead of antennas.

We further enrich the throughput measurements by varying the noise power levels in the RFSim mode and the signal level in the SMA cable mode since noise power level could further impact the network channel quality. Specifically, in the RFSim mode, in addition to the default perfect channel, we enable channel modeling and set the noise power to -1dBm, -5dBm, and -10dBm at the UE side. In the SMA cable mode,

we configure different reduced amplitude levels (30dB, 40dB and 50dB) of the incoming signal at the UE side by connecting two 10dB attenuators and one 30dB attenuator in series.

2) Measurement Settings: We configure the RAN network to operate on 5G band n78 (3.5 GHz) with 40 MHz of spectrum. Since our measurements focus on 5G RAN network, the test is performed in the OAI `phy-test` setup which enables the communication between the gNB and UE without the need of a core network. We use `iperf` as the workload generator to send UDP traffic. The UDP bandwidth at the gNB is configured to 1 Mbit/sec, which is restricted by the processing power of the two host machines. We vary the MCS index from 0 to 28, with MCS indices 29, 30, and 31 reserved by 3GPP.

B. Measurement Results

Fig. 1(c) summarizes the throughput results as a function of the MCS index under various connectivity settings. Intuitively, one would expect the throughput to increase monotonically with higher MCS indices, as higher modulation schemes and coding rates typically result in higher data rates and channel efficiency. However, upon observation, it is apparent that only the results of RFSim with default channel and SMA cable connection with the 50dB attenuator meet this expectation. All other results show fluctuations as the MCS indices increase.

For example, in the RFSim mode, when an extremely high level of noise (-1dBm) is added to the simulated channel, the network throughput is very low (<80 Kbits/sec, values enlarged by 10 times for improved visibility in Fig. 1(c)). The data also show significant fluctuation when the MCS indices are small (from 0 to 9). When we further increase the MCS index, the throughput directly drops to 0 (values omitted in Fig. 1(c)) due to extremely high packet loss rate. A similar trend can be observed in all the other curves, but the occurrences of fluctuation vary, and the throughput fluctuation is delayed under better channel conditions.

C. Discussions

Based on the throughput results obtained in our 5G RAN testbed, we have the following observation.

Observation 1. *Only in the case of a high-quality channel the throughput monotonically increases with the increase of MCS index. Under worsening channel conditions, the throughput can decrease with higher MCS indices, and significant fluctuations may occur.*

The fluctuation in throughput with increased MCS index is mainly caused by the following reason. The communication channel between the UE and the gNB is composed of a set of subbands, and the channel quality may vary across these subbands. However, according to the PHY specification of 5G NR [13], the UE must select and use the same MCS index on all the allocated subbands. When the MCS index is increased, the UE may fail to decode the received signals on subbands with poor channel quality, resulting in decreased throughput. However, on subbands with good channel quality, the UE can achieve higher data rates, leading to improved throughput. These opposite trends in throughput changes on different subbands result in overall throughput fluctuations across the entire bandwidth as the MCS index increases.

Regarding Observation 1, there are two points worth noting. First, the OTA mode measurement on our testbed is conducted in a line-of-sight (LOS) indoor lab environment with no moving objects nor significant interference/noise sources (see Fig. 1(a)), and the throughput results show fluctuation when MCS index is larger than 9. When considering industrial 5G RAN networks which are typically deployed in much harsher environments, the channel quality can be much worse and the fluctuation in network performance can be more significant for different MCS indices.

Secondly, in our testbed, the gNB is connected with only one UE which has access to the entire network bandwidth resource. The gNB only needs to determine the MCS index used by the UE to achieve better performance, e.g., higher throughput. However, if multiple UEs are connected to the RAN network, it becomes more challenging to determine the proper MCS index for each UE, given that the bandwidth is shared by all the UEs and the subband allocation for each UE also needs to be determined. Therefore, based on the experimental results and the above discussion, we make the following statement to motivate our work.

Statement. *The selection of MCS index for each UE is crucial in determining the performance of 5G RAN networks, and presents a challenge that requires judicious investigation in the design of scheduling mechanisms.*

III. System Model and Problem Statement

We now present the 5G NR-based network model and formulate the 5G real-time flow scheduling problem.

A. Network and Traffic Model

We consider a single-cell DL 5G RAN system where one gNB serves a set of N UEs, both of which are equipped with a MIMO antenna panel, with M_T antennas at the gNB and M_R antennas at each UE $u_i \in \mathcal{U}$ ($i \in [1, N]$) ($M_T > M_R$).

1) OFDMA Resource Grid: In OFDMA-based 5G NR, as shown in Fig. 2, network resource is organized as a resource grid that spans in both time and frequency domains. In the time domain, time is equally slotted into transmission time intervals (TTIs) each of which consists of 14 OFDM symbols [37]. In the frequency domain, bandwidth of the operating channel is divided into a number of uniform subbands, and each subband is denoted as a resource block (RB). That is, within each TTI, there is a set of RBs $\mathcal{B}^+ = \{b | b \in [1, 2, \ldots, B]\}$ that can be allocated to the UEs for transmissions, where B represents the total number of RBs in the frequency domain.

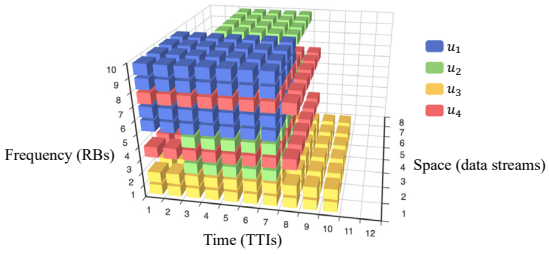

Fig. 2. Resource grid in 5G NR. Each block represents a basic time-frequency-space scheduling unit for UEs.

2) Traffic Model: Communication in industrial applications is typically characterised by two attributes, *periodicity* and *determinism*, which together specify periodic traffic flows with stringent timing requirements [1]. To simplify the notation, we assume that each UE $u_i \in \mathcal{U}$ $(i \in [1, N])$ receives one transmission flow, denoted as $f_i \in \mathcal{F}$ $(i \in [1, N])$, from the gNB periodically[1]. Each f_i is associated with a tuple $\langle P_i, D_i, C_i \rangle$. P_i and D_i denote the period and deadline of f_i (in unit of TTIs), respectively, and we assume $D_i \leq P_i$. C_i denotes the payload size (in bits) which is the amount of information carried in each instance of f_i. The k-th instance of flow f_i is referred to as packet $p_{i,k}$. Its release time and absolute deadline are denoted as $r_{i,k}$ and $d_{i,k}$, respectively.

3) MCS Model: The objective of the real-time flow scheduling problem in 5G NR is to allocate RBs to individual flows in the flow set $\mathcal{F}$ while satisfying their timing requirements. A number of 5G specific issues must be considered in formulating the scheduling problem. Besides RB allocation, the scheduler also needs to select a proper MCS for each UE in each TTI [13]. As discussed in Section II, a larger MCS index generally leads to a higher UE data rate. However, the maximum data rate that can be achieved on one RB depends on the channel condition between UE and gNB. If the channel condition on this RB is poor but a high MCS is used, data may not be successfully received by the UE.

Channel conditions can vary in both time (across different TTIs, i.e., time-selective fading) and frequency (across different RBs, i.e., frequency-selective fading). Variation of channel condition in the time domain is mainly determined by motion effects, e.g., UEs installed on moving objects and obstacles moving between UEs and the gNB [38]. The channel condition is reported from individual UEs to the gNB through the CQI either periodically or aperiodically which is configured by the Radio Resource Control (RRC) message(s). In the frequency domain, channel attenuation, which suffers from severe fading effects (e.g., reflective obstacles such as machines), is non-negligible. Therefore, the channel condition between each UE and the gNB varies on different RBs in the frequency domain.

We denote $\mathcal{M} = \{0, 2, \ldots, 28\}$ as the set of 29 available MCS indices defined in [13]. Let q_i^b be the maximum MCS index that can be used by UE u_i on RB $b \in \mathcal{B}^+$ so that

[1]The model can be generalized by treating multiple flows of one UE as multiple UEs.

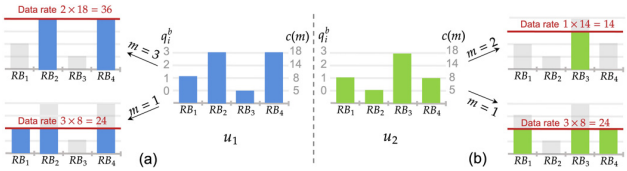

Fig. 3. A MCS selection example for u_1 and u_2 on 4 RBs. The colored blocks represent RBs with data transmissions under MCS $m \leq q_i^b$. The shaded blocks represent RBs with no data transmission under MCS $m > q_i^b$.

data carried on b can be successfully received, and we have $0 \leq q_i^b \leq 28$. q_i^b is determined according to the subband CQI submitted by UE u_i on RB b. Let $c(m)$ be the modulation and coding rate on an RB under MCS m, and $a_i^{b,m}$ be the achievable data rate on RB b for UE u_i under MCS m. If $m \leq q_i^b$, the data can be successfully transmitted and the achievable data rate is $c(m)$. Otherwise, i.e., $m > q_i^b$, the transmission fails with data rate being 0 [13][2]. That is,

$$a_i^{b,m} = \begin{cases} c(m) & \text{if } m \leq q_i^b, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

According to the 5G NR PHY specification [13], although each UE can be allocated with multiple RBs in one TTI, it must select and use the same MCS index on all the allocated RBs. For example, suppose there are 4 RBs in the frequency domain (i.e., $B = 4$) and the channel conditions (i.e., q_i^b) on the 4 RBs for two UEs u_1 and u_2 are shown in Fig. 3. If we select MCS $m = 1$ for u_1, then data carried on RB_1, RB_2 and RB_4 can be successfully transmitted, and the total data rate is $3 \times 8 = 24$. If we select a higher MCS $m = 3$, a higher data rate can be achieved on RB_2 and RB_4, i.e., $2 \times 18 = 36$. However, a higher MCS index is not always leading to a higher data rate, according to the measurement results in our motivational experiments. For instance, setting a higher MCS $m = 2$ for u_2 leads to a lower data rate 14 compared to data rate 24 that can be achieved by setting $m = 1$.

4) MU-MIMO: In 5G NR, MU-MIMO [41] is supported to improve the spectral efficiency by spatially multiplexing multiple UEs. That is, the gNB communicates with multiple UEs on the same RB within one TTI. For example, in Fig. 2, UEs u_1 and u_2 are simultaneously transmitted on RB_{10} within TTIs $[3, 7]$. On each RB $b \in \mathcal{B}^+$, the scheduler selects a set of UEs for MIMO transmission, where $s_i^b(t) \in \{0, 1\}$ is a binary variable indicating if RB $b \in \mathcal{B}^+$ is scheduled by the gNB for UE u_i in TTI t. Further, each UE u_i can simultaneously transmit multiple data streams (also called "layers" in specification [42]) on each RB in one TTI. For example, in Fig. 2, 6 data streams are transmitted by u_3 on RB_1 and RB_2 in each TTI within TTIs $[1, 10]$.

To reduce the control signaling overhead and signal processing complexity, 5G NR requires that each UE must have the

[2]$c(m)$ and q_i^b can be determined through channel estimation using existing methods (e.g., [39], [40]) and looking up MCS index table from [13], which are assumed as given to the formulated scheduling problem in this paper.

same number of data streams across all RBs allocated to it in each TTI. Let $y_i(t)$ be the number of data streams transmitted by user u_i in TTI t. If the number of antennas for each UE is M_R, we have $y_i(t) \leq M_R$. Similarly, the total number of data streams on each RB b for MIMO transmissions cannot exceed the number of antennas at the gNB, M_T, i.e.,

$$\sum_{u_i \in \mathcal{U}} s_i^b(t) \cdot y_i(t) \leq M_T, \forall b \in \mathcal{B}^+. \tag{2}$$

In a practical MU-MIMO system, propagation channels among different UEs are spatially correlated and the theoretical multi-antenna gains cannot always be obtained. Channel correlations can generate interference among UEs due to the channel fading of multipath propagation between the gNB and UEs with diverse spatial transmission directions. To ensure the transmission performance of MU-MIMO, the scheduler needs to perform user selection to avoid data streams from the UEs with high interference to be transmitted on the same RB which may deteriorate the achievable data rate [43]. In this paper, we use $w(u_i, u_j) \in \{0, 1\}$ to denote whether two UEs u_i and u_j can transmit on the same RB simultaneously according to their interference level, where $w(u_i, u_j) = 0$ indicates that u_i and u_j cannot transmit simultaneously[3]. We assume that $w(u_i, u_j)$ for each pair of UEs, which can be derived by existing work (e.g. [21]), is known at the gNB. We denote θ_i as the achievable data rate flag for user u_i in a MU-MIMO system, where θ_i captures the impact on the data rate of u_i from other UEs that simultaneously transmit on the same RB.

$$\theta_i = \begin{cases} 0 & \text{if there exists } u_j \text{ transmitting on the} \\ & \text{same RB with } u_i \text{ and } w(u_i, u_j) = 0, \\ 1 & \text{otherwise.} \end{cases} \tag{3}$$

Transmitted data amount. The total amount of data transmitted to u_i in t across all its allocated RBs, denoted as $R_i(t)$, can be calculated by $R_i(t) = \sum_{b \in \mathcal{B}^+} \left(s_i^b(t) \cdot a_i^{b, m_i(t)} \cdot y_i(t) \cdot \theta_i \right)$, where $m_i(t)$ denotes the selected MCS index for u_i in t. Then, the total amount of data transmitted to UE u_i in packet $p_{i,k}$ equals to $R_{i,k} = \sum_{t \in [r_{i,k}, d_{i,k})} R_i(t)$.

B. 5G Real-Time Flow Scheduling Problem

Based on the above system model, the task of the network scheduler at the gNB is to generate a *schedule* that determines the resource allocation for all the UEs.

Definition 1 (Schedule)**.** *A schedule specifies the following resource allocation decisions for each UE u_i in each TTI t.*

- *The RBs allocated to u_i, i.e., $\{s_i^b(t) | b \in \mathcal{B}^+\}$;*
- *The selected MCS index for u_i, i.e., $m_i(t)$;*
- *The number of data streams transmitted to u_i, i.e., $y_i(t)$.*

[3]In practical MU-MIMO systems, two UEs may always transmit on the same RB simultaneously, and the data rate degradation of each UE depends on the level of interference. In this paper, we simplify the MAC layer flow model by allowing (prohibiting) two UEs from transmitting on the same RB simultaneously if the gNB determines their interference level as low (high).

Given the definition of a schedule, we aim to solve the following real-time flow scheduling problem in 5G NR.

Problem P: Given the UE set $\mathcal{U}$, flow set $\mathcal{F}$, the modulation and coding rate $c(m)$ on any RB, the maximum MCS index q_i^b usable by UE u_i on any RB b, and the interference information $w(u_i, u_j)$ between UEs, determine a feasible schedule (if exists) that satisfies the deadlines of all the packets released by the flows in $\mathcal{F}$.

In the following sections, we outline our 5G NR scheduling framework. We first assume a stable channel condition within each hyperperiod H (i.e., the least common multiple of all the flow periods), where q_i^b for each UE is updated once every H TTIs. We then extend the system model and scheduling method to account for dynamic channel conditions, where q_i^b for each UE is updated when the channel condition changes.

IV. Overall Scheduling Framework

Problem **P** is NP-hard and this can be proved by reducing the set packing problem [44] to a special case of the SISO (single-input-single-output) version of Problem **P**. The details of the proof is omitted here due to the space limit.

The solution space of Problem **P** is also extremely large. Specifically, the gNB needs to allocate B RBs among N UEs and assign each UE an optimal MCS (among 29 possible indices) in each TTI. On each RB b, the number of possible combinations of simultaneously transmitted flows equals to $\binom{N}{1} + \binom{N}{2} + \cdots + \binom{N}{M_T}$ and the number of possible combinations of data streams is $(M_R)^N$. This gives a total number of $\left(N^B \cdot (29 \cdot M_R)^N \cdot \left[\binom{N}{1} + \binom{N}{2} + \cdots + \binom{N}{M_T}\right]\right)^H$ possibilities in the solution space. Given the complexity and the large solution space of Problem **P**, we design a **two-phase** scheduling framework, named 5G-TPS, to judiciously reduce the search space following a set of insights. The key principle of 5G-TPS is to maximize the channel efficiency for all the UEs such that all the flows can meet their timing requirements.

At the highest level, we adopt a channel condition aware approach to generate a schedule for all the flows $f_i \in \mathcal{F}$. Specifically, when the network channel condition is stable within each hyperperiod, we apply the same RB allocation to each flow in all its scheduled TTIs. This approach has one distinct advantage. That is, it reduces the overhead for communicating Downlink Control Information (DCI). Note that, individual RB allocations across different TTIs require multiple DCI messages each of which specifies one RB allocation with individual TTI information and the selected MCS index. This incurs large control resource overhead which, in turn, reduces the amount of network resources allocated to PDSCH (Physical Downlink Shared Channel) for transmitting actual data. On the other hand, if an RB allocation is 'satisfactory' to all the flows in $\mathcal{F}$ in one TTI, it is not necessary to make any adjustment on the RB allocations in other TTIs when the maximum usable MCS index q_i^b for each UE is not changed (Theorem 1 below demonstrates this observation). Therefore,

using the same RB allocation for each flow in all its scheduled TTIs is sufficient.

Based on the general approach outlined above, in **Phase 1** of 5G-TPS, we aim to find a feasible RB allocation across all the TTIs in the hyperperiod to satisfy all the flows' deadlines. If Phase 1 fails, i.e., an RB allocation satisfying the deadlines of all the flows cannot be found, **Phase 2** is activated to adjust the RB allocations based on the output of Phase 1. Specifically, the redundant RBs allocated to certain flows in the unused TTIs together with some unallocated RBs will be judiciously assigned to the unschedulable flows to meet their deadlines.

If the channel condition changes within each hyperperiod in the form of q_i^b update from each affected UE, we perform **schedule adjustment** among different flows, in terms of MCS index re-selection and RB allocation adjustments. Details of the schedule adjustment will be discussed in Section VII.

V. RB Allocation in Phase 1 of 5G-TPS

In this section, we describe Phase 1 of 5G-TPS by focusing on two questions: i) what is a feasible RB allocation that is satisfactory to all the flows? and ii) how to find such a feasible RB allocation? Before answering these two questions, we first determine the setting on the number of data streams for each UE u_i (i.e., $y_i(t)$) in the spatial domain.

In MU-MIMO 5G NR, each UE is equipped with M_R antennas and can receive at most M_R concurrent data streams on each RB. Thus, a larger data rate can be achieved by each flow $f_i \in \mathcal{F}$ on each RB if multiple data streams are configured by its corresponding UE u_i. Then, more data of packet $p_{i,k}$ can be transmitted within each TTI and the transmission of $p_{i,k}$ can complete earlier. In addition, more data streams configured for each flow leads to less number of UEs simultaneously receiving in each TTI, since the total number of data streams from all the flows cannot exceed the number of antennas equipped on the gNB (i.e., M_T). This results in less transmission interference among different UEs caused by channel correlation and thus a higher channel efficiency can be achieved. Based on the above observations, **we fix $y_i(t) = M_R$ for each f_i in 5G-TPS**[4].

A. Flow Set Schedulability

By setting $y_i(t) = M_R$ for each flow f_i, the lemma below defines the feasible RB allocation for each flow (i.e., answering the first question in Phase 1 design).

Lemma 1. *If M_R data streams are transmitted by u_i in each TTI and the amount of transmitted data per data stream is larger than or equal to $\left\lceil \frac{C_i}{D_i \cdot M_R} \right\rceil$, flow f_i is schedulable, i.e., satisfies the deadline.*

[4]For many 5G commercial products, the number of antennas equipped on the gNB is an integral multiple of the number of antennas on UE, e.g., 64×64 MIMO supported by Nokia AirScale and 4×4 MIMO supported by Quectel RM510Q. Therefore, we argue that $M_T \mod M_R = 0$, and setting $y_j(t) = M_R$ does not result in any resource waste due to the insufficient number of available data streams, which is less than M_R.

The proof of Lemma 1 is straightforward and thus omitted. According to Lemma 1, an RB allocation for flow f_i, denoted as $\mathcal{B}_i \subseteq \mathcal{B}^+$, is defined as feasible if the total amount of data transmitted on RBs $b \in \mathcal{B}_i$ in one TTI is larger than or equal to $\left\lceil \frac{C_i}{D_i \cdot M_R} \right\rceil$. Based on Lemma 1, we give the theorem below to define the schedulability of flow set $\mathcal{F}$.

Theorem 1. *If an RB allocation for all the flows $f_i \in \mathcal{F}$ ($i \in [1, N]$) in one TTI, denoted as $\Theta = \{\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_N\}$, satisfies the following three constraints, flow set $\mathcal{F}$ is schedulable.*

Constraint 1. *Each RB allocation $\mathcal{B}_i \in \Theta$ is feasible for f_i according to Lemma 1.*

Constraint 2. *For any two UEs u_i and u_j with $w(u_i, u_j) = 0$, no common RB exists in the RB allocations $\mathcal{B}_i$ and $\mathcal{B}_j$.*

Constraint 3. *Each RB $b \in \mathcal{B}^+$ can be allocated to at most $\frac{M_T}{M_R}$ flows in each TTI.*

Proof Sketch. According to Lemma 1, Constraint 1 guarantees that each flow f_i is schedulable with allocated RBs in $\mathcal{B}_i$. Constraint 2 guarantees that no transmission interference occurs between any two UEs. Constraint 3 satisfies the limit on the number of antennas equipped on the gNB (i.e., Eq. (2)). □

To find a feasible RB allocation Θ for $\mathcal{F}$ (i.e., answering the second question in Phase 1 design), we first determine a feasible *RB allocation candidate set* for each flow f_i, denoted as $\{\mathcal{B}_i^*\}$ (i.e., each $\mathcal{B}_i \in \{\mathcal{B}_i^*\}$ satisfies Constraint 1). We then formulate an RB allocation selection problem to select one RB allocation $\mathcal{B}_i$ for each flow f_i from its candidate set $\{\mathcal{B}_i^*\}$ such that Constraints 2&3 in Theorem 1 are satisfied. Below we elaborate these two steps.

B. RB Allocation Candidate Set Generation

Over the entire network bandwidth, there exists a large number of RB allocations for each flow (e.g., $\sum_{x=1}^{B} \binom{B}{x} = 2^B - 1$ if all RB allocations are feasible) satisfying Constraint 1, creating an extensive search space for the RB allocation selection problem. To improve the search efficiency, we aim to generate a small RB allocation candidate set $\{\mathcal{B}_i^*\}$ for each f_i, and only include in it the most promising RB allocations (instead of solving in one shot the RB allocation selection problem based on all feasible RB allocations of each flow).

Generating $\{\mathcal{B}_i^*\}$ out of all feasible RB allocations is a challenging task. To tackle this, we explore the relationship between two critical factors: i) the number of RBs allocated to each flow and ii) the achievable data rate of each flow. The number of RBs allocated to a flow impacts the available RB allocations for other flows in $\mathcal{F}$ since the limited number of RBs over the entire bandwidth are shared by all the flows. On the other hand, the achievable data rate of each flow determines its own schedulability. According to Lemma 1, the feasible RB allocations must have achievable data rate larger than or equal to $\left\lceil \frac{C_i}{D_i \cdot M_R} \right\rceil$. Furthermore, higher data rates for each UE u_i at a given number of RBs are more desirable since each packet of f_i can transmit the required C_i amount of data using less

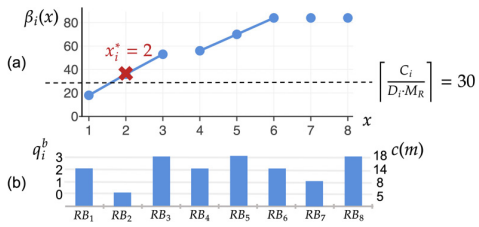

Fig. 4. (a) An example of the highest data rate function $\beta_i(x)$. The dashed line represents the required data rate. (b) The channel conditions of user u_i on a network with 8 RBs.

TTIs where the RBs within the unused TTIs can be utilized by other flows to complete their data transmissions in Phase 2. Thus, we introduce the *highest data rate function* for each flow to aid us identify the flow's feasible RB allocations.

Definition 2 (Highest data rate function $\beta_i(x)$)**.** *$\beta_i(x)$ is the highest data rate that can be achieved by f_i per data stream if x number of RBs in $\mathcal{B}^+$ are allocated to f_i.*

Obtaining $\beta_i(x)$ is to find x RBs with the 'best' channel conditions such that flow f_i can achieve the highest data rate. $\beta_i(x)$ can be calculated by traversing the maximum usable MCS index q_i^b on all the RBs $b \in \mathcal{B}^+$ in a descending order, and storing the highest data rate achieved using each MCS index q_i^b. This process ends until we find at least x RBs with the maximum usable MCS index equal to the current q_i^b value. For example, Fig. 4 shows $\beta_i(x)$ for flow f_i on a network with 8 RBs. When calculating $\beta_i(4)$, it starts from $q_i^b = 3$, and only three RBs are with the maximum usable MCS index equal to 3, thus the achievable data rate is $3 \times 18 = 54$. We proceed with $q_i^b = 2$, and 6 RBs (larger than 4) are with the maximum usable MCS index larger than or equal to 2 where the achievable data rate is $4 \times 14 = 56$. Thus, we have $\beta_i(4) = 56$.

As depicted in Fig. 4, all the RB allocations resulting $\beta_i(x) \geq \left\lceil \frac{C_i}{D_i \cdot M_R} \right\rceil (x \leq B)$ can form an RB allocation candidate set $\{\mathcal{B}_i^*\}$ for flow f_i. However, the size of this set is still large due to two reasons: i) large networks may have a significant number of RBs (i.e., a large B), and ii) each $\beta_i(x)$ value can correspond to multiple RB allocations. For example, in Fig. 4, $\beta_i(2) = 36$ and there exist three RB allocations with $\alpha_i(\{3,5\}) = \alpha_i(\{3,8\}) = \alpha_i(\{5,8\}) = 36$, where $\alpha_i(\mathcal{B})$ denotes the highest achievable data rate with RB allocation $\mathcal{B}$.

Therefore, we outline our findings through several important lemmas below, which provide guidelines on reducing the set of considered feasible RB allocations, i.e., generating the RB allocation candidate set $\{\mathcal{B}_i^*\}$ for each flow f_i.

Lemma 2. *The highest data rate function $\beta_i(x)$ is segmented by piecewise linear functions of x, denoted as $\beta_i(x) = \{\beta_{i,j}(x) = c(\zeta_j) \cdot x | x \in \{x_j, x_j + 1, \ldots, x_j'\}\}$, where $c(\zeta_j)$ denotes the achievable data rate under MCS index ζ_j. For any two linear segments $\beta_{i,j}(x)$ and $\beta_{i,h}(x)$, if $x_j < x_h$, we have $c(\zeta_j) > c(\zeta_h)$.*

Proof Sketch. When the number of RBs having a larger usable MCS ζ_j is greater than the current value x, ζ_j can be selected as the MCS index and $\beta_i(x)$ increases with the same increment $c(\zeta_j)$. When x is greater than the number of RBs having ζ_j, a smaller MCS index ζ_h has to be selected and the slope of $\beta_i(x)$ reduces to $c(\zeta_h)$[5]. □

Lemma 2 indicates that when x increases, if $\beta_i(x)$ transfers to another linear function, the maximum usable MCS index to achieve $\beta_i(x)$ decreases. This leads to a set of RBs on each of which flow f_i transmits under a MCS index lower than q_i^b. For example, in Fig. 4, $\beta_i(4) = \alpha_i(\{1,3,4,5\})$ and the MCS index used is $m = 2$, i.e., $c(2) \times 4 = 14 \times 4 = 56$. However, the maximum usable MCS index on RB_3 and RB_5 is 3. That is, the channel efficiency achieved on these two RBs decreases. Therefore, Lemma 2 motivates us to only select the values of x within the first linear function of $\beta_i(x)$ that satisfies Constraint 1, i.e., $\beta_i(x) \geq \left\lceil \frac{C_i}{D_i \cdot M_R} \right\rceil$. For instance, in Fig. 4, the set of considered number of allocated RBs in the candidate set is reduced from $x \in \{2,3,\ldots,8\}$ to $x \in \{2,3\}$.

Lemma 3. *Consider the number of allocated RBs x following a same linear function, i.e., $\beta_i(x) = c(\zeta_j) \cdot x, x \in \{x_j, x_j + 1, \ldots, x_{j'}\}$. For any RB allocation $\mathcal{B}'(x_j + 1 \leq |\mathcal{B}'| = x' \leq x_{j'})$ such that $\alpha_i(\mathcal{B}') = \beta_i(x')$, there must exists at least one RB allocation $\mathcal{B}^o(|\mathcal{B}^o| = x_j)$ such that $\alpha_i(\mathcal{B}^o) = \beta_i(x_j)$ and $\mathcal{B}^o$ is a subset of $\mathcal{B}'$, i.e., $\mathcal{B}^o \subset \mathcal{B}'$.*

Proof. Since x' and x_j follow a same linear function, according to Lemma 2, $\beta_i(x') = c(q_i^{b_j}) \cdot x'$ and $\beta_i(x_j) = c(q_i^{b_j}) \cdot x_j$. Thus, for any RB allocation $\mathcal{B}'$ such that $\alpha_i(\mathcal{B}') = c(q_i^{b_j}) \cdot x'$, there exist $x' > x_j$ RBs with the maximum usable MCS higher than or equal to $q_i^{b_j}$. Then, we can have an RB allocation $\mathcal{B}^o \subset \mathcal{B}'$ by selecting arbitrary x_j RBs in $\mathcal{B}'$ such that $\alpha_i(\mathcal{B}^o) = c(q_i^{b_j}) \cdot x_j = \beta_i(x_j)$. □

Lemma 3 indicates that any RB allocation resulting in $\beta_i(x')$ is a superset of an RB allocation resulting in $\beta_i(x_j)$, if x' and x_j $(x' > x_j)$ follow a same linear function. Then, if all the RB allocations resulting in $\beta_i(x_j)$ cannot satisfy Constraint 2 in the RB allocation selection process, the RB allocations resulting in $\beta_i(x')$ cannot either. This motivates us to only consider the minimum value of x within a linear function of $\beta_i(x)$ that satisfies Constraint 1. For instance, in Fig. 4, the set of considered number of allocated RBs is further reduced from $x \in \{2,3\}$ to $x = 2$.

Based on Lemma 2&3, we can determine the RB allocation candidate set $\{\mathcal{B}_i^*\}$ for each flow f_i as follow.

RB allocation candidate set determination. We determine the number of RBs allocated to flow f_i, denoted as x_i^*, as the minimum x satisfying $\beta_i(x) \geq \left\lceil \frac{C_i}{D_i \cdot M_R} \right\rceil$, and any RB allocation $\mathcal{B}$ resulting $\alpha_i(\mathcal{B}) = \beta_i(x_i^*)$ is included in the RB allocation candidate set $\{\mathcal{B}_i^*\}$. For example, in Fig. 4, $x_i^* = 2$ and $\{\mathcal{B}_i^*\} = \{\mathcal{B} | \alpha_i(\mathcal{B}) = \beta_i(2)\} = \{\{3,5\}, \{5,8\}, \{3,8\}\}$.

[5]Due to the page limit, the formal proof is included in the supplemental material which is available from the Program Chair upon request.

C. RB Allocation Selection

After an RB allocation candidate set $\{\mathcal{B}_i^*\}$ is generated for each f_i, we need to select one RB allocation $\mathcal{B}_i \in \{\mathcal{B}_i^*\}$ for each f_i such that Constraints 2&3 are satisfied. If a feasible RB allocation cannot be found for any flow in $\mathcal{F}$, Phase 2 is triggered to adjust the RB allocation based on the output of Phase 1. Therefore, we formulate an RB allocation selection problem **P1** as an optimization problem to maximize the number of schedulable flows in Phase 1.

Problem P1. Given the RB allocation candidate set $\{\mathcal{B}_i^*\}$ for each flow $f_i \in \mathcal{F}$, determine a schedulable flow set $\mathcal{F}_1$ where i) each flow $f_i \in \mathcal{F}_1$ is assigned with an RB allocation $\mathcal{B}_i \in \{\mathcal{B}_i^*\}$, ii) Constraints 2&3 are satisfied, and iii) the size of $\mathcal{F}_1$ is maximized.

Problem **P1** is NP-hard since it is equivalent to the set packing problem and any heuristic designed for solving a set packing problem can be applied to solve **P1** (e.g., [45]).

VI. RB Allocation in Phase 2 of 5G-TPS

In Phase 1, each flow f_i is allocated with the same set of RBs in all the TTIs within the hyperperiod. Although this allocation reduces overhead and simplifies the scheduling problem since only RB allocations in the frequency domain need to be considered, it may allocate unnecessary RBs for certain flows in the time domain (i.e., in certain TTIs) with unused data streams in the spatial domain. This waste of resources may lead to unschedulable flows. To solve this issue, this section presents a solution to Problem **P** in Phase 2 to satisfy the timing requirements of the unschedulable flows by using those RBs with available data streams in certain TTIs.

A. Remaining RB Set

The remaining RBs in the output of Phase 1 include unallocated RBs and unused RBs. The former are the set of RBs that are not allocated to any flows in Phase 1. The latter is the set of RBs that are allocated to UEs but not used by the corresponding flows in certain TTIs.

Remaining RBs in the time domain. As described in Section V-B, if the achieved data rate of flow f_i in some TTIs is larger than the requirement based on Lemma 1 (i.e., $\beta_i(x^*) > \left\lceil \frac{C_i}{D_i \cdot M_R} \right\rceil$), f_i only needs $\left\lceil \frac{C_i}{\beta_i(x_i^*) \cdot M_R} \right\rceil$ TTIs to complete the transmission of each released packet, where the RBs in the rest $P_i - \left\lceil \frac{C_i}{\beta_i(x_i^*) \cdot M_R} \right\rceil$ TTIs are not used.

Remaining RBs in the spatial domain. If the total number of data streams transmitted by the flows on an RB in one TTI is less than the number of antennas on the gNB (i.e., M_T), this RB can be used as a remaining RB with a set of available data streams. As an example, the white blocks in Fig. 5(a) and Fig. 5(b) represent the remaining RBs in the time and spatial domain, respectively.

B. Phase 2 Overview

In Phase 2, we use the remaining RBs to generate a feasible schedule for each unschedulable flow $f_j \in \mathcal{F}_2 = \mathcal{F} - \mathcal{F}_1$,

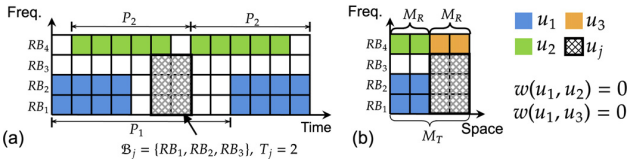

Fig. 5. Illustration of Phase 2 scheduling for flow f_j using the remaining RBs in the time and spatial domain in (a) and (b), respectively. The colored blocks represent the RB allocation for scheduled flows in Phase 1. The white blocks represent the remaining RBs. The patterned blocks represent the feasible RB allocation for f_j.

where $\mathcal{F}_1$ is the set of flows feasibly scheduled in Phase 1. Specifically, for f_j we determine the RB allocation and MCS index in the frequency domain, the scheduled TTIs in the time domain, and the number of data streams in the spatial domain.

In the time domain, we schedule each packet of f_j in a consecutive set of TTIs to reduce control overhead, given that each DCI message only carries 4 bits for the time domain resource assignment (the start TTI index and the number of TTIs) according to 3GPP [46]. In the spatial domain, we follow the setting of $y_j = M_R$ to reduce the transmission interference among UEs and improve the channel efficiency.

Thus, in Phase 2, we assign each packet $p_{j,k}$ of flow $f_j \in \mathcal{F}_2$ with a feasible schedule specifying RB allocation $\mathcal{B}_j$ and TTI duration, denoted as $S_{j,k} = [t_{j,k}, t_{j,k} + T_j)$, where $t_{j,k}$ and T_j are the start TTI and length of the consecutive TTIs, respectively. That is, all the packets released by f_j share the same $\{\mathcal{B}_j, T_j\}$ configuration with individual start TTIs $t_{j,k}$. Theorem 2 below specifies the schedulability of flow set $\mathcal{F}_2$. The proof is omitted due to the similarity to that of Theorem 1.

Theorem 2. *If the schedule of each flow $f_j \in \mathcal{F}_2$, denoted as $\{\mathcal{B}_j, S_j\}$ where $S_j = \{S_{j,k|k=1,2,\ldots}\}$, satisfies the following constraints, flow set $\mathcal{F}_2$ is schedulable.*

Constraint 4. *For any packet $p_{j,k}$, $t_{j,k} \geq r_{j,k}$, $t_{j,k} + T_j \leq r_{j,k} + D_j$, and $\alpha(\mathcal{B}_j) \cdot T_j \geq \left\lceil \frac{C_i}{M_R} \right\rceil$.*

Constraint 5. *If $\mathcal{B}_j$ shares a common RB in a TTI with any $\mathcal{B}_i$ ($f_i \in \mathcal{F}$), $w(u_i, u_j) = 1$.*

Constraint 6. *For any RB $b \in \mathcal{B}_j$ and TTI in $S_{j,k}$, the number of flows share the same RB is not larger than M_T/M_R.*

To summarize, since both the RB allocation in the frequency domain and the TTI configuration in the time domain are considered for flows in Phase 2, we generate the schedule for each flow $f_j \in \mathcal{F}_2$ in an iterative fashion to avoid combinatorial explosion among RB allocations for all the flows in different TTIs. In each iteration, given the set of remaining RBs within TTIs $[1, H]$, we i) determine $\{\mathcal{B}_j, S_j\}$ for flow f_j with the highest utilization (i.e., C_j/P_j) in $\mathcal{F}_2$ since f_j typically requires more RBs in each TTI than the other flows, and ii) update the set of remaining RBs.

C. Feasible Schedule Generation

In this section, we describe how to generate the feasible schedule $\{\mathcal{B}_j, S_j\}$ for flow f_j using the remaining RBs. We

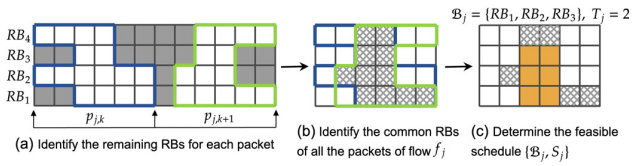

Fig. 6. Illustration of the feasible schedule generation for flow f_j. (a) The white blocks represent the remaining RBs of packet $p_{j,k}$ and $p_{j,k+1}$. (b) The patterned blocks represent the common RBs of the two packets. (c) The orange blocks represent the determined feasible schedule for f_j.

first give the problem formulation.

Problem P2. Given the set of remaining RBs, the specification of flow f_j, determine a feasible schedule $\{\mathcal{B}_j, S_j\}$ satisfying all the constraints in Theorem 2.

A feasible schedule $\{\mathcal{B}_j, S_j\}$ specifies a 'rectangle' with $|\mathcal{B}_j|$ RBs in the frequency domain[6] and T_j TTIs in the time domain. According to Theorem 2, we need to guarantee two requirements: (i) the rectangle must exist within the period of each packet $p_{j,k}$, and (ii) the amount of data transmitted in the rectangle must be larger than or equal to C_j.

To satisfy requirement (i), we traverse the set of remaining RBs usable by all the packets of f_j and identify the common RBs (i.e., RBs in the same relative TTIs in the period of each packet, see Fig. 6(a) and Fig. 6(b)). For requirement (ii), generating a feasible schedule satisfying Constraint 4 is equivalent to finding a rectangle of area C_j, where the length equals to T_j and the height equals to $c(m)\cdot|\mathcal{B}_j|$ (see Fig. 6(c)). Since the data rate of individual RBs is different, the width of the rectangle is not only determined by the number of RBs, $|\mathcal{B}_j|$, but also the set of allocated RBs and the corresponding MCS index. This problem with non-identical RBs is a variation of the largest empty rectangle problem where many efficient algorithms can be applied, e.g., [47], [48].

If a feasible schedule can be generated for each flow $f_j \in \mathcal{F}_2$, Problem **P** is solved, i.e., all the flows in $\mathcal{F}$ are schedulable. Note that, 5G-TPS aims to generate a schedule for flow set $\mathcal{F}$ in an efficient and effective manner. For this reason, some solutions of Problem **P** may be pruned from the search space (e.g., some other RB allocations leading to higher $\beta_i(x)$ with a larger x in Phase 1). Therefore, our solution only provides a sufficient schedulability condition for flow set $\mathcal{F}$.

VII. Dynamic Schedule Adjustment

In industrial 5G NR, the channel condition between a UE and the gNB can vary over time caused by moving obstacles, multipath propagation and interference from other devices, etc. In this section, we generalize the system model to consider channel dynamics and present a dynamic schedule adjustment method based on the two-phase design of 5G-TPS.

As shown in Fig. 7, when the network channel condition is stable, the feasible schedule of each UE is carried out through the DCI messages on PDCCH (Physical Downlink Control Channel). Upon any channel condition change being measured by UE u_e, it sends an updated CQI report to the gNB on PUCCH (Physical Uplink Control Channel) to specify the new q_e^b values on certain RBs $b \in \mathcal{B}^+$. To respond to the channel condition change, the gNB adjusts the schedule(s) for u_e and other UEs if needed, and transmits the updated schedule via the subsequent DCI messages.

[6]Here we refer to a logical rectangle since the RB allocation in the frequency domain may not be consecutive.

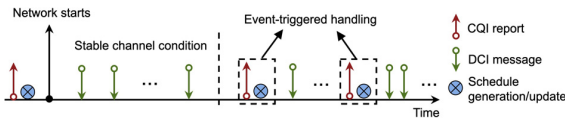

Fig. 7. The network execution model for handling dynamic channel.

The gNB may receive multiple updated CQI reports from different UEs within a short time interval. In this case, the gNB just recomputes the schedules for all the affected UEs. Therefore, we follow an event-triggered mechanism to perform schedule adjustment at the gNB, when the channel condition changes for a particular UE u_e, with the aim to satisfy the timing requirements of all the flows. The schedule adjustment problem can be defined as follows.

Problem P3. Given the updated set $\{q_e^b|(b \in \mathcal{B}^+)\}$ for UE u_e, the schedule of each flow $f_i \in \mathcal{F}$, determine the schedule adjustment to meet the deadlines of all the flows $f_i \in \mathcal{F}$.

The channel condition change of u_e can be classified into three cases.

Case 1: $\forall b \in \mathcal{B}_e, a_e^{b,m} = c(m)$. In this case, the maximum usable MCS q_e^b on each allocated RB b is still larger than or equal to the selected MCS m. That is, the achieved data rate on $\mathcal{B}_e$ of flow f_e in each scheduled TTI still satisfies the data rate requirement according to Lemma 1. Thus, the gNB does not need to adjust the schedule for flow f_e.

Case 2: $\exists b \in \mathcal{B}_e, a_e^{b,m} < c(m)$, but $\exists m', \alpha(\mathcal{B}_e) \geq \left\lceil \frac{C_e}{T_e \cdot M_R} \right\rceil$. In this case, the maximum usable MCS q_e^b on certain allocated RB(s) b is smaller than MCS m, and the achieved data rate on each of these RB(s) drops to 0 according to the MCS model. However, another MCS index m' can be used to achieve a data rate higher than or equal to the requirement, i.e., $\alpha(\mathcal{B}_e) \geq \left\lceil \frac{C_e}{T_e \cdot M_R} \right\rceil$. Thus, the gNB only updates the MCS index for u_e.

Case 3: $\alpha(\mathcal{B}_e) < \left\lceil \frac{C_e}{T_e \cdot M_R} \right\rceil$. In this case, the amount of data that can be transmitted by each packet of f_e is less than its payload size according to its current schedule, thus the schedule needs to be adjusted.

To handle Case 3, we leverage the remaining RBs to schedule f_e using Phase 2. i.e., solving Problem **P2**. If using the remaining RBs cannot satisfy the timing requirement of f_e for the channel condition change, we re-generate the schedule for all the flows $f_i \in \mathcal{F}$ using Phase 1 and Phase 2[7]. If Phase 1 and Phase 2 fail, we deem that flow f_e is unschedulable after channel condition change.

[7]Adjusting only a subset of flows does not save much control overhead since each flow without adjustment still needs DCI messages specifying its subsequent schedules.

App.	Flow specifications					Experiment results										
	Service area (m^2)	Msg. size (bytes)	Period (ms)	Deadline (ms)	# of UEs	# of packets	Latency distribution									
							Stable channel					Dynamic channel				
							1 (ms)	2 (ms)	3 (ms)	4 (ms)	5 (ms)	1 (ms)	2 (ms)	3 (ms)	4 (ms)	5 (ms)	∞
ES	900	250	8	4	16	80000	29590	10410				17937	6218		15845		
S	900	1024	10	5	20	80000		4	4	29964	10028		4	4	17922	21882	188
VC	1800	100	10	5	10	40000	19912	64	24			11910	316	42	4	7732	
HF	100	50	2	1	4	80000	40000					40000					

(a)

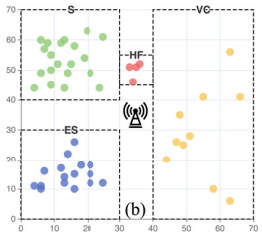

Fig. 8. Case study. (a) The flow specifications in the considered use case and the summarized latency results. (b) UE locations.

VIII. Performance Evaluation

In this section, we perform a simulation-based case study and conduct extensive simulation experiments to evaluate the proposed 5G-TPS framework. Although we established a real-world 5G RAN testbed, as described in Sec. II, conducting 5G-TPS performance evaluations on the testbed is not feasible for two reasons. From the hardware aspect, the current 5G testbed only consists of one gNB and one UE, and the high cost of USRP devices makes it difficult to create a large-scale testbed for experimental evaluation of a large set of UEs. From the software aspect, the OAI 5G project currently only supports wideband CQI report, where the UE reports one single q_i for the entire bandwidth. However, the scheduling mechanism proposed in this work is based on 5G subband CQI report (i.e., q_i^b per RB), which is not on the OAI's roadmap in the near future and the implementation of subband CQI is non-trivial and out of the scope of this work.

A. Case Study

Setup. We perform a case study using the specifications of the mobile operation panel use case provided by 3GPP (Table A.2.4.1A-1 in [1]). The use case consists of four applications: 1) Emergency Stop (ES) for connectivity availability, 2) Safety (S) data stream, 3) Visualization of Control (VC), and 4) Haptic Feedback (HF) data stream. The flow specifications and UE locations are provided in Fig. 8. We consider a 20 MHz bandwidth network consisting of 100 RBs. The average q_i^b of each UE is configured according to the path loss effect based on its distance to the gNB. We set $M_T = 8$ and $M_R = 4$.

We let the network run for $40s$, i.e., 1000 hyperperiods. In the first 500 hyperperiods, we apply a stable channel condition where q_i^b for each UE is randomly updated once every H TTIs. In the rest 500 hyperperiods, we assume dynamic channel conditions, where q_i^b for each flow is frequently changed every two packets. Fig. 8(a) summarizes the latency experienced by all the packets released from individual applications.

Results. In the stable channel condition, although the packets from a same application may have different latency, they can always be transmitted within their deadlines. In the dynamic channel condition, the latency distribution of each flow is more scattered since more schedule adjustments are performed for each flow. On the other hand, flow S encounters a fraction of deadline misses (188 packets out of 80000 packets, denoted as ∞), which happens when a feasible schedule cannot be generated for flow S. However, most of the packets from all the applications still meet their deadlines.

B. Experiment Setup

To evaluate the performance of 5G-TPS under various network settings, we generate a large number of random synthetic flow sets. To speed up the simulation which involves many network nodes, we do not perform computational PHY processing of the air interface but focus on the MAC layer scheduler evaluation.

1) Variables: The used variables include the number of RBs B, the number of flows N, and the normalized flow set utilization $U^* \in (0, 1]$ where $U^* = \sum_{f_i \in \mathcal{F}} \frac{C_i}{P_i(B \cdot c(|\mathcal{M}|) \cdot M_T)}$. Here U^* captures the flow set workload on one RB per data stream with the maximum modulation and coding rate $c(|\mathcal{M}|)$. $U^* = 1$ means that the flow set can be schedulable only if the maximum MCS $|\mathcal{M}|$ can be used by each UE on all RBs $b \in \mathcal{B}^+$ (i.e., under ideal channel condition) and the number of antennas of each UE equals to M_T.

2) Metrics: We use Schedulability Ratio (SR) as the metric for performance evaluation under the stable channel condition setting. SR is defined as the ratio of feasible flow sets to all the generated flow sets. In the dynamic channel condition, we use the number of Deadline Missed flows (DM) to evaluate the performance of 5G-TPS in schedule adjustment.

3) Compared Methods: We compare the performance of 5G-TPS with the following scheduling methods.

SMT: The Satisfiability Modulo Theory-based exact solution (the SMT specifications are omitted due to page limit).

MUST: A 5G NR scheduler aiming at maximizing the number of packets delivered within the deadlines [49]. MUST relies on a greedy approach to assigning the most urgent packets with RBs of the highest data rate.

CA: A channel condition-aware response time analysis for 5G networks under fixed-priority scheduling [26]. CA is based on an over-simplified resource model where the entire bandwidth is treated as one single RB. A generalized version considering multiple RBs, denoted as **CA-Ext**, is also implemented[8].

[8]Extending the response time analysis of CA in networks consisting of multiple RBs is non-trivial. Here, we directly run the fixed priority scheduler which provides a safe upper bound on the SR/DM achieved by CA-Ext.

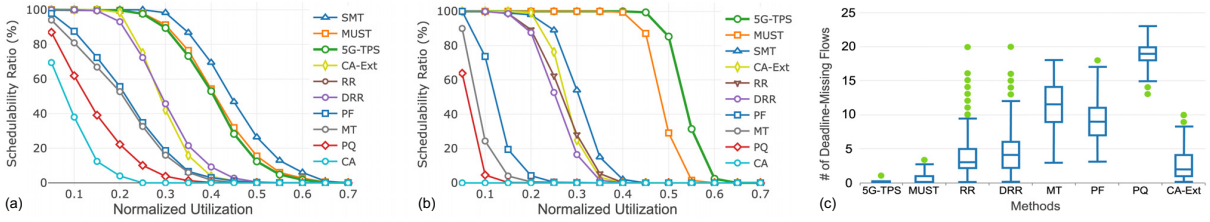

Fig. 9. Evaluation results. (a) SR comparisons in small scale networks. (b) SR comparisons in large networks. (c) DM comparisons in large networks.

RR, **MT**, and **PF**: The built-in flow schedulers (i.e., round-robin, maximum CQI, and proportional fair) in OAI 5G [50].

DRR and **PQ**: Two extended flow schedulers (i.e., deficit RR and priority queue) based on the built-in schedulers [51].

C. Experiment Results

1) Stable Channel Condition: In the first set of experiments, we compare the SRs of all the methods by varying the normalized utilization U^* under stable channel conditions. Due to the runtime limitation suffered by the SMT-based solution, we make the performance comparison under two network settings: (1) an extremely small scale network with 3 UEs, 8 RBs, $M_R = 2$ and $M_T = 4$, and (2) a more practical large scale network with $N = 25$, $B = 50$, $M_R = 4$ and $M_T = 16$. Further, we set a $3000s$ timeout limit for the SMT approach in large scale networks to prevent it from spending a long time for finding a result.

Small scale networks. Fig. 9(a) shows the SR as a function of U^* in small scale networks. Each point represents the average value of 5000 trials. The results show that the SRs of all the methods decrease with the increasing of U^* and SMT dominates others as an exact solution. The SR gap between SMT and 5G-TPS is small (4.63% on average) which validates the effectiveness of 5G-TPS. On the other hand, 5G-TPS significantly outperforms most of the other methods (e.g., 15.44% higher than DRR on average) and shows almost the same SR with MUST (0.79% lower on average). However, MUST degrades significantly when the network scales to the normal size (i.e., $N = 25$, $B = 50$ as shown in the next section). The SR of CA is very low (only 8.84% on average) while the extended version CA-Ext shows a much higher SR (38.32% on average). This demonstrates the limitation of the over-simplified resource model used in [26].

Large scale networks. Fig 9(b) shows the SR as a function of U^* in large scale networks. We can see that the SRs of SMT, PF, MT, PQ and CA drop significantly compared to those in the small scale networks (17.97% lower on average), and the SRs of CA-Ext, RR and DRR drop slightly (3.66% lower on average). The degradation of SMT is mainly because it fails in most cases under the timeout limit. For example, when $U^* = 0.6$ and SR $= 0\%$, only 7.3% flow sets are determined by SMT as unschedulable while all other failures are caused by timeout. The SR drops of the other methods demonstrate that they cannot properly perform resource allocation for many UEs, even with more available network resources.

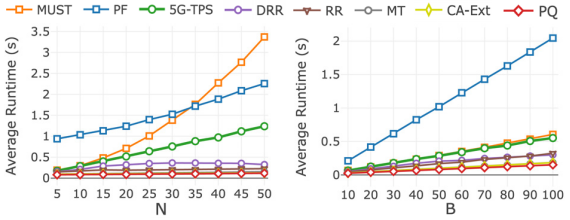

Fig. 10. Runtime comparisons with varying N and B.

On the other hand, the SRs of 5G-TPS and MUST increase, where the SR increase of 5G-TPS (18.36% higher on average) is much larger than that of MUST (9.84% higher on average). This demonstrates that 5G-TPS can better utilize the network resources to accommodate a large amount of real-time flows.

2) Dynamic Channel Condition: In the second set of experiments, we evaluate all the methods in large scale networks with dynamic channel conditions. We randomly generate one flow set with $U^* = 0.4$ and run continuously for 200 hyperperiods. q_i^b of each UE u_i is randomly updated once within each hyperperiod. In this experiment we do not evaluate SMT and CA because the high overhead of SMT hinders it from being applied for on-line schedule adjustment and the performance of CA is dominated by CA-Ext according to the results in the previous experiments.

Fig. 9(c) shows the DM distributions of all the methods and the result of each method represents the DM statistics in one hyperperiod. We can observe that all the methods suffer from deadline misses for certain flows. However, 5G-TPS outperforms all the others in terms of much lower DM where only one flow misses its deadline in one hyperperiod (i.e., the outlier 1). MUST satisfies the deadlines of most flows with an average DM of 0.47 where at most 3 flows miss their deadlines in one hyperperiod. However, MUST results missing deadline flows in 66 hyperperiods. The other methods suffer from higher DMs, especially for MT, PF and PQ, where flows miss their deadlines in all the hyperperiods.

3) Runtime: Since 5G-TPS may rerun both Phase 1 and Phase 2 online in response to channel condition changes, we evaluate the runtime of 5G-TPS to validate its online adoption. Our primary focus is on comparing the runtime trends across different methods, rather than the absolute runtime values since our experiments are conducted in Python on a single-

core processor which is not comparable to the performance of commercial gNBs (e.g., Atom P5900 with up to 24 cores in Nokia AirScale Radios). In addition, SMT suffers from extremely high runtime (e.g., $>$ 2 hours when $N > 5$ or $B > 10$), and thus cannot be used for runtime reconfiguration. Such high overhead also hindered us from including SMT in the comparison. Fig. 10 presents the average runtime of each method, with N and B as variables.

From the results, we can observe a linear increase in the runtime of most methods with the increases of N and B. However, the runtime of MUST experiences an explosive growth with the increase of N, possibly due to its per-TTI and packet-based scheduling design. In contrast, 5G-TPS consumes less time compared to MUST (38.12% lower on average) and the built-in scheduler PF (57.29% lower on average), demonstrating its efficiency.

IX. Related Work

A. Ultra-Reliable Low-Latency Communications

Supporting ultra-reliable low-latency communications (URLLC) traffic is one of the key revolutionary novelties of 5G NR. URLLC targets ensuring extremely low latency in the order of $1ms$ and providing high reliability of 99.999% for flows with timing requirements. Some existing works on URLLC focus on the lower layer functionality (e.g., designing new robust MCSs [52]–[54], HARQ enhancement [55], [56] and power management [57]–[59], etc.). On the other hand, many recent works study the scheduling problems for URLLC in 5G NR which can be classified into two categories.

In the first category, the problems of joint scheduling for coexisting URLLC and enhanced mobile broadband (eMBB) traffic (e.g., web, video) are widely studied [60]–[62]. However, all these works mainly focus on maximizing the performance of eMBB traffic in the presence of sparse URLLC traffic. They all assume that the timing requirements of URLLC traffic can be satisfied by immediately scheduling them upon arrival and preempting the ongoing eMBB traffic. They do not consider massive URLLC scenarios where a large number of URLLC flows contend for the network resource.

In the second category, several scheduling algorithms and resource allocation methods based on Proportional Fair (PF) are proposed for URLLC traffic [23], [63]–[65]. However, all of these works do not address whether flows can be delivered before their deadlines. Alternatively, some approaches have been proposed to satisfy the timing requirements of URLLC traffic [49], [66], [67]. In [67] and [66], the authors simplify the scheduling problem by considering either Frequency Division Multiple Access (FDMA), which is inefficient in the time domain, or Time Division Multiple Access (TDMA), which is inefficient in the frequency domain. Instead, [49] considers OFDMA-based 5G networks, similar to our work. However, they simply ignore the data rate impact of MCS selection on different RBs and do not consider MIMO networks.

B. MU-MIMO

There have been many active studies on designing traffic schedulers for MIMO cellular networks (e.g., [21], [68]–[72]). In [70], [71], the authors propose a set of scheduling algorithms for MU-MIMO users to optimize the system throughput. [69] studies the user selection and resource allocation problem with the goal of maximizing the user sum rate in the context of 802.11ax. In [68], the authors address the problem of frequency domain packet scheduling in MIMO LTE but do not consider MCS selection. Thus, none of the above works can jointly optimize the configuration of RBs and MCS for MU-MIMO users. In [21], the authors present a 5G scheduler mCore to maximize the network throughput with joint optimization of RB allocation and MCS selection to MU-MIMO users. A deep Q-network-based joint adaptive scheduling algorithm of MCS and space division multiplexing in 5G massive MIMO-OFDM is proposed in [72] to maximize the network throughput. However, none of these works provides guarantees on satisfying the deadline requirements of real-time flows in 5G NR.

C. MCS Selection

A significant amount of research efforts have been made on adaptive MCS in wireless communication systems. For instance, an aggressive MCS selection method is proposed in [73] to maximize the system throughput in an OFDM system. [74] gives a comprehensive analysis of MCS selection in MIMO-OFDM systems which takes into account hybrid ARQ (HARQ) operation. [75] proposes an optimal MCS selection criterion for maximizing user throughput in cellular networks. All these traditional research on MCS selection mainly study the mapping between the channel quality and MCS level, which is assumed given in this work. To further improve the system performance in terms of throughput, many works focus on cross-layer scheduling algorithm design together with MCS selection, e.g., [24], [76], [77]. However, all these works do not consider hard real-time requirements of 5G flows.

X. Conclusion

In this paper, we leverage a real-world 5G RAN testbed to benchmark the DL throughput with varying MCS indices and formulate the real-time flow scheduling problem in industrial 5G NR, which features per-flow real-time schedulability guarantees through time-frequency-space resource allocation. We propose a two-phase scheduling framework, namely 5G-TPS, to construct a feasible schedule with deadline guarantees for all the flows in 5G NR and enable online schedule adjustment for flows upon dynamic channel condition changes. Our extensive experimental results demonstrate the superior performance of 5G-TPS when compared to other state-of-the-art scheduling approaches in 5G NR, in terms of schedulability ratio, under both stable and dynamic channel conditions.

As for the future work, we will explore learning-based dynamic RB allocation strategies and adjustable OFDMA numerology with varied TTI sizes.

XI. Acknowledgement

The work is supported in part by the National Science Foundation (NSF) Grant CNS-1932480, CNS-2008463 and CCF-2028875.

References

[1] 3GPP, "Service requirements for cyber-physical control applications in vertical domains," *3rd Generation Partnership Project (3GPP), Technical Specification (TS) 22.104*, 06 2021, version 18.1.0.

[2] A. Aijaz, "Private 5g: The future of industrial wireless," *IEEE Industrial Electronics Magazine*, vol. 14, no. 4, pp. 136–145, 2020.

[3] J. Jasperneite, M. Schumacher, and K. Weber, "Limits of increasing the performance of industrial ethernet protocols," in *EFTA*. IEEE, 2007, pp. 17–24.

[4] G. Wang, T. Zhang, C. Xue, J. Wang, M. Nixon, and S. Han, "Time-sensitive networking for industrial automation: Challenges, opportunities, and directions," *arXiv preprint arXiv:2306.03691*, 2023.

[5] S. Petersen and S. Carlsen, "Wirelesshart versus isa100. 11a: The format war hits the factory floor," *IEEE Industrial Electronics Magazine*, vol. 5, no. 4, pp. 23–34, 2011.

[6] R. Steigmann and J. Endresen, "Introduction to wisa: Wisa-wireless interface for sensors and actuators," *White paper, ABB*, 2006.

[7] D. Dujovne, T. Watteyne, X. Vilajosana, and P. Thubert, "6tisch: deterministic ip-enabled industrial internet (of things)," *IEEE Communications Magazine*, vol. 52, no. 12, pp. 36–41, 2014.

[8] J. Wang, T. Zhang, D. Shen, X. S. Hu, and S. Han, "Harp: Hierarchical resource partitioning in dynamic industrial wireless networks," in *2022 IEEE 42nd International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2022, pp. 1029–1039.

[9] B. Bellalta, "Ieee 802.11 ax: High-efficiency wlans," *IEEE Wireless Communications*, vol. 23, no. 1, pp. 38–46, 2016.

[10] 3GPP, "Technical specification group services and system aspects," *3rd Generation Partnership Project (3GPP), Technical Report (TR) 21.915*, 2019, version 15.0.0.

[11] B. Makki, K. Chitti, A. Behravan, and M.-S. Alouini, "A survey of noma: Current status and open research challenges," *IEEE Open Journal of the Communications Society*, vol. 1, pp. 179–189, 2020.

[12] X. Lin, J. Li, R. Baldemair, J.-F. T. Cheng, S. Parkvall, D. C. Larsson, H. Koorapaty, M. Frenne, S. Falahati, A. Grovlen *et al.*, "5g new radio: Unveiling the essentials of the next generation wireless access technology," *IEEE Communications Standards Magazine*, vol. 3, no. 3, pp. 30–37, 2019.

[13] 3GPP, "Nr; physical layer procedures for data," *3rd Generation Partnership Project (3GPP), Technical Specification (TS) 38.214*, 2018, version 15.0.0.

[14] A. Saifullah, Y. Xu, C. Lu, and Y. Chen, "End-to-end communication delay analysis in industrial wireless networks," *IEEE Transactions on Computers*, vol. 64, no. 5, pp. 1361–1374, 2014.

[15] V. P. Modekurthy, A. Saifullah, and S. Madria, "Distributedhart: A distributed real-time scheduling system for wirelesshart networks," in *RTAS*. IEEE, 2019, pp. 216–227.

[16] T. Zhang, T. Gong, Z. Yun, S. Han, Q. Deng, and X. S. Hu, "Fd-pas: A fully distributed packet scheduling framework for handling disturbances in real-time wireless networks," in *RTAS*. IEEE, 2018, pp. 1–12.

[17] J. Wang, T. Zhang, D. Shen, X. S. Hu, and S. Han, "Apas: An adaptive partition-based scheduling framework for 6tisch networks," in *RTAS*. IEEE, 2021, pp. 320–332.

[18] Y.-H. Wei, Q. Leng, S. Han, A. K. Mok, W. Zhang, and M. Tomizuka, "Rt-wifi: Real-time high-speed communication protocol for wireless cyber-physical control applications," in *RTSS*. IEEE, 2013, pp. 140–149.

[19] A. Mamane, M. E. Ghazi, G.-R. Barb, and M. Oteșteanu, "5g heterogeneous networks: an overview on radio resource management scheduling schemes," in *2019 7th Mediterranean congress of telecommunications (CMT)*. IEEE, 2019, pp. 1–5.

[20] J. Wang, Y. Liu, S. Niu, and H. Song, "Reinforcement learning optimized throughput for 5g enhanced swarm uas networking," in *ICC 2021-IEEE International Conference on Communications*. IEEE, 2021, pp. 1–6.

[21] Y. Chen, Y. Wu, Y. T. Hou, and W. Lou, "mcore: Achieving sub-millisecond scheduling for 5g mu-mimo systems," in *INFOCOM*. IEEE, 2021, pp. 1–10.

[22] Y. Chen, Y. T. Hou, W. Lou, J. H. Reed, and S. Kompella, "M 3: A sub-millisecond scheduler for multi-cell mimo networks under c-ran architecture," in *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*. IEEE, 2022, pp. 130–139.

[23] 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. 72 253–72 262, 2018.

[24] C. Li, Y. Huang, Y. Chen, B. Jalaian, Y. T. Hou, and W. Lou, "Kronos: A 5g scheduler for aoi minimization under dynamic channel conditions," in *ICDCS*. IEEE, 2019, pp. 1466–1475.

[25] C. Li, Y. Huang, S. Li, Y. Chen, B. A. Jalaian, Y. T. Hou, W. Lou, J. H. Reed, and S. Kompella, "Minimizing aoi in a 5g-based iot network under varying channel conditions," *IEEE Internet of Things Journal*, vol. 8, no. 19, pp. 14 543–14 558, 2021.

[26] A. Nota, S. Saidi, D. Overbeck, F. Kurtz, and C. Wietfeld, "Context-based latency guarantees considering channel degradation in 5g network slicing," in *2022 IEEE Real-Time Systems Symposium (RTSS)*. IEEE, 2022, pp. 253–265.

[27] ——, "Providing response times guarantees for mixed-criticality network slicing in 5g," in *2022 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2022, pp. 552–555.

[28] A. Shashin, A. Belogaev, A. Krasilov, and E. Khorov, "Adaptive parameters selection for uplink grant-free urllc transmission in 5g systems," *Computer Networks*, vol. 222, p. 109527, 2023.

[29] Y. Pan, R. Mahfouzi, S. Samii, P. Eles, and Z. Peng, "Resource optimization with 5g configured grant scheduling for real-time applications," in *2023 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2023, pp. 1–2.

[30] T. Zhang, X. S. Hu, and S. Han, "Contention-free configured grant scheduling for 5g urllc traffic," in *Proceedings of the 60th ACM/IEEE Design Automation Conference*, 2023.

[31] A. Saifullah, Y. Xu, C. Lu, and Y. Chen, "Real-time scheduling for wirelesshart networks," in *2010 31st IEEE Real-Time Systems Symposium*. IEEE, 2010, pp. 150–159.

[32] V. P. Modekurthy, A. Saifullah, and S. Madria, "A distributed real-time scheduling system for industrial wireless networks," *ACM Transactions on Embedded Computing Systems (TECS)*, vol. 20, no. 5, pp. 1–28, 2021.

[33] D. Shen, T. Zhang, J. Wang, Q. Deng, S. Han, and X. S. Hu, "Distributed successive packet scheduling for multi-channel real-time wireless networks," in *2022 IEEE 28th International Conference on Embedded and Real-Time Computing Systems and Applications (RTCSA)*. IEEE, 2022, pp. 71–80.

[34] S. S. Nakkina, S. Balijepalli, and C. R. Murthy, "Performance benchmarking of the 5g nr phy on the oai codebase and usrp hardware," in *WSA 2021; 25th International ITG Workshop on Smart Antennas*. VDE, 2021, pp. 1–6.

[35] Y. Huang, Y. T. Hou, and W. Lou, "Deluxe: A dl-based link adaptation for urllc/embb multiplexing in 5g nr," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 1, pp. 143–162, 2021.

[36] N. Nikaein, M. K. Marina, S. Manickam, A. Dawson, R. Knopp, and C. Bonnet, "Openairinterface: A flexible platform for 5g research," *ACM SIGCOMM Computer Communication Review*, vol. 44, no. 5, pp. 33–38, 2014.

[37] 3GPP, "Nr; physical channels and modulation," *3rd Generation Partnership Project (3GPP), Technical Specification (TS) 38.211*, vol. 9, 2018.

[38] M. Düngen, T. Hansen, R. Croonenbroeck, R. Kays, B. Holfeld, D. Wieruch, P. W. Berenguer, V. Jungnickel, D. Block, U. Meier *et al.*, "Channel measurement campaigns for wireless industrial automation," *at-Automatisierungstechnik*, vol. 67, no. 1, pp. 7–28, 2019.

[39] A. Le Ha, T. Van Chien, T. H. Nguyen, W. Choi *et al.*, "Deep learning-aided 5g channel estimation," in *2021 15th international conference on ubiquitous information management and communication (IMCOM)*. IEEE, 2021, pp. 1–7.

[40] H. A. Le, T. Van Chien, T. H. Nguyen, H. Choo, and V. D. Nguyen, "Machine learning-based 5g-and-beyond channel estimation for mimo-ofdm communication systems," *Sensors*, vol. 21, no. 14, p. 4861, 2021.

[41] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive mimo for next generation wireless systems," *IEEE communications magazine*, vol. 52, no. 2, pp. 186–195, 2014.

[42] 3GPP, "Nr; multiplexing and channel coding," *3rd Generation Partner-*

ship Project (3GPP), Technical Specification (TS) 38.212, 2019, version 15.7.0.

[43] L. Sanguinetti, E. Björnson, and J. Hoydis, "Toward massive mimo 2.0: Understanding spatial correlation, interference suppression, and pilot contamination," *IEEE Transactions on Communications*, vol. 68, no. 1, pp. 232–257, 2019.

[44] X. Gandibleux, X. Delorme, and V. T'Kindt, "An ant colony optimisation algorithm for the set packing problem," in *International Workshop on Ant Colony Optimization and Swarm Intelligence*. Springer, 2004, pp. 49–60.

[45] B. Chandra and M. M. Halldórsson, "Greedy local improvement and weighted set packing approximation," *Journal of Algorithms*, vol. 39, no. 2, pp. 223–240, 2001.

[46] 3GPP, "Nr; physical layer procedures for control," *3rd Generation Partnership Project (3GPP), Technical Specification (TS) 38.213*, 2020, version 15.10.0.

[47] A. Aggarwal and S. Suri, "Fast algorithms for computing the largest empty rectangle," in *Proceedings of the third annual symposium on Computational geometry*, 1987, pp. 278–290.

[48] M. Orlowski, "A new algorithm for the largest empty rectangle problem," *Algorithmica*, vol. 5, pp. 65–73, 1990.

[49] E. Khorov, A. Krasilov, I. Selnitskiy, and I. F. Akyildiz, "A framework to maximize the capacity of 5g systems for ultra-reliable low-latency communications," *IEEE Transactions on Mobile Computing*, vol. 20, no. 6, pp. 2111–2123, 2020.

[50] Openairinterface 5g ran project. [Online]. Available: https://gitlab.eurecom.fr/oai/openairinterface5g/-/tree/develop

[51] R.-M. Ursu, A. Papa, and W. Kellerer, "Experimental evaluation of downlink scheduling algorithms using openairinterface," in *2022 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2022, pp. 84–89.

[52] C. Yue, M. Shirvanimoghaddam, B. Vucetic, and Y. Li, "Channel coding and decoding schemes for urllc," *Ultra-Reliable and Low-Latency Communications (URLLC) Theory and Practice: Advances in 5G and Beyond*, pp. 119–168, 2023.

[53] Y. Samarawickrama, Á. A. de Medeiros, and V. Cionca, "Joint optimization of ofdm and channel coding for urllc in industrial channels," *IEEE Transactions on Industrial Informatics*, 2022.

[54] C. Yue, V. Miloslavskaya, M. Shirvanimoghaddam, B. Vucetic, and Y. Li, "Efficient decoders for short block length codes in 6g urllc," *IEEE Communications Magazine*, vol. 61, no. 4, pp. 84–90, 2023.

[55] F. Nadeem, M. Shirvanimoghaddam, Y. Li, and B. Vucetic, "Nonorthogonal harq for urllc: Design and analysis," *IEEE Internet of Things Journal*, vol. 8, no. 24, pp. 17 596–17 610, 2021.

[56] S. AlMarshed, D. Triantafyllopoulou, and K. Moessner, "Deep learning-based estimator for fast harq feedback in urllc," in *2021 IEEE 32nd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*. IEEE, 2021, pp. 642–647.

[57] C. Yin, R. Zhang, Y. Li, Y. Ruan, T. Li, T. Tao, and D. Li, "Packet re-management-based c-noma for urllc: From the perspective of power consumption," *IEEE Communications Letters*, vol. 26, no. 3, pp. 682–686, 2021.

[58] M. Ganjalizadeh, H. S. Ghadikolaei, A. Azari, A. Alabbasi, and M. Petrova, "Saving energy and spectrum in enabling urllc services: A scalable rl solution," *IEEE Transactions on Industrial Informatics*, 2023.

[59] C. Yin, R. Zhang, Y. Li, Y. Ruan, T. Tao, and D. Li, "Power consumption minimization for packet re-management based c-noma in urllc: Cooperation in the second phase of relaying," *IEEE Transactions on Wireless Communications*, 2022.

[60] H. Yin, L. Zhang, and S. Roy, "Multiplexing urllc traffic within embb services in 5g nr: Fair scheduling," *IEEE Transactions on Communications*, vol. 69, no. 2, pp. 1080–1093, 2020.

[61] D. Shen, T. Zhang, J. Wang, Q. Deng, S. Han, and X. S. Hu, "Qos guaranteed resource allocation for coexisting embb and urllc traffic in 5g industrial networks," in *2022 IEEE 28th International Conference on Embedded and Real-Time Computing Systems and Applications (RTCSA)*. IEEE, 2022, pp. 81–90.

[62] M. Darabi, V. Jamali, L. Lampe, and R. Schober, "Hybrid puncturing and superposition scheme for joint scheduling of urllc and embb traffic," *IEEE Communications Letters*, vol. 26, no. 5, pp. 1081–1085, 2022.

[63] A. A. Esswie and K. I. Pedersen, "Multi-user preemptive scheduling for critical low latency communications in 5g networks," in *2018 IEEE Symposium on Computers and Communications (ISCC)*. IEEE, 2018, pp. 00 136–00 141.

[64] A. Akhtar and H. Arslan, "Downlink resource allocation and packet scheduling in multi-numerology wireless systems," in *2018 IEEE wireless communications and networking conference workshops (WCNCW)*. IEEE, 2018, pp. 362–367.

[65] M. Almekhlafi, M. A. Arfaoui, C. Assi, and A. Ghrayeb, "Superposition-based urllc traffic scheduling in 5g and beyond wireless networks," *IEEE Transactions on Communications*, vol. 70, no. 9, pp. 6295–6309, 2022.

[66] A. Destounis, G. S. Paschos, J. Arnau, and M. Kountouris, "Scheduling urllc users with reliable latency guarantees," in *2018 16th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*. IEEE, 2018, pp. 1–8.

[67] A. Pratap, R. Gupta, V. S. S. Nadendla, and S. K. Das, "On maximizing task throughput in iot-enabled 5g networks under latency and bandwidth constraints," in *2019 IEEE International Conference on Smart Computing (SMARTCOMP)*. IEEE, 2019, pp. 217–224.

[68] S.-B. Lee, I. Pefkianakis, S. Choudhury, S. Xu, and S. Lu, "Exploiting spatial, frequency, and multiuser diversity in 3gpp lte cellular networks," *IEEE Transactions on Mobile Computing*, vol. 11, no. 11, pp. 1652–1665, 2011.

[69] K. Wang and K. Psounis, "Scheduling and resource allocation in 802.11 ax," in *INFOCOM*. IEEE, 2018, pp. 279–287.

[70] K. Ko, J. Lee, and W. Shin, "Joint power allocation and scheduling techniques for ber minimization in multiuser mimo systems," *IEEE Access*, vol. 9, pp. 66 675–66 686, 2021.

[71] J. Kang and W. Yu, "Scheduling versus contention for massive random access in massive mimo systems," *IEEE Transactions on Communications*, vol. 70, no. 9, pp. 5811–5824, 2022.

[72] Y. Liao, Z. Yang, Z. Yin, and X. Shen, "Dqn-based adaptive mcs and sdm for 5g massive mimo-ofdm downlink," *IEEE Communications Letters*, 2022.

[73] A. Talukdar, P. Sartori, M. Cudak, B. Classon, and Y. Blankenship, "Aggressive modulation/coding scheme selection for maximizing system throughput in a multi-carrier system," in *2005 IEEE 61st Vehicular Technology Conference*, vol. 5. IEEE, 2005, pp. 3038–3042.

[74] S. Liu, X. Zhang, and W. Wang, "Analysis of modulation and coding scheme selection in mimo-ofdm systems," in *2006 First International Conference on Communications and Electronics*. IEEE, 2006, pp. 240–245.

[75] D. Kim, B. C. Jung, H. Lee, D. K. Sung, and H. Yoon, "Optimal modulation and coding scheme selection in cellular networks with hybrid-arq error control," *IEEE transactions on wireless communications*, vol. 7, no. 12, pp. 5195–5201, 2008.

[76] H. Song, K. J. Kim, J. Guo, P. V. Orlik, and K. Parsons, "Semi-persistent scheduling scheme for low-latency and high-reliability transmissions in private 5g networks," in *2022 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, 2022, pp. 651–656.

[77] L. N. Dinh, M. Maman, and E. C. Strinati, "Proactive resource scheduling for 5g and beyond ultra-reliable low latency communications," in *2022 IEEE 95th Vehicular Technology Conference:(VTC2022-Spring)*. IEEE, 2022, pp. 1–5.