# 5G-TPS: A Two-Phase Real-Time Scheduling and Adaptation Framework for 5G Radio Access Networks

Tianyu Zhang, *Member, IEEE*, Jiachen Wang, X. Sharon Hu, *Fellow, IEEE*, and Song Han, *Member, IEEE*

***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, wide coverage, ultra-low latency, and massive connectivity. Despite its great potential, 5G NR also brings significant complexity in scheduling data flows to meet their hard real-time requirements in industrial applications. In this paper, we first leverage a 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 guarantee through time-frequency resource allocation. We propose a novel two-phase scheduling framework, named 5G-TPS, to construct a 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. For large-scale multi-cell 5G industrial systems with cloud radio access network (C-RAN) architecture, we further introduce a user association algorithm respecting the real-time requirements of individual user equipment (UEs). Extensive experimental studies 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.**

***Index Terms*—Real-time 5G, flow scheduling, MCS selection, industrial Internet-of-Things.**

## 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 (e.g., process control, automotive, and aerospace manufacturing). Many of these industrial applications (e.g., use cases specified by 3GPP [1] including wired-to-wireless link replacement, 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].

Industrial wireless networks enable more flexible network configurations [3] and reduce cabling costs compared to their wired counterparts [4] (e.g., Time-Sensitive Networking [5]). However, existing industrial wireless solutions (e.g., ISA100.11, WirelessHART, and 6TiSCH [6], [7], [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) have received growing attention in industrial applications due to their low deployment cost. However, 802.11-based protocols operate in an unlicensed spectrum and may suffer severe and unexpected interference from other co-existing networks (e.g., WirelessHART and Bluetooth). Further, the limitations on coverage, mobility, and outdoor deployment make them only suitable for indoor industrial applications [2].

The industrial connectivity landscape is changing with the emergence of 5G New Radio (NR) systems. 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 offer complete control to configure every aspect of the network (e.g., schedule, resource allocation, and security).

To achieve ultra-high-speed communication with stringent timing and reliability requirements, several enabling technologies are supported in 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, and shorter transmission time intervals (TTIs) compared to 4G LTE are applied to reduce latency. In addition, 5G NR provides robust modulation and coding schemes (MCS), which determine the user's data rate on individual frequency bands according to the per-band channel quality indicator (CQI), namely the subband CQI report.

Despite the great potential of 5G NR in industrial applications, it brings high complexity due to the large design space of the flow scheduler to meet the real-time requirements of industrial data flows. Specifically, the scheduler needs to i) allocate resource blocks (RBs) appropriately in the frequency domain

Received 18 February 2025; revised 7 July 2025; accepted 13 August 2025. Date of publication 19 August 2025; date of current version 3 December 2025. This work was supported in part by NSF under Grant CNS-2008463 and in part by the Investment in Strategic Priorities funding from the Univeristy of Iowa. Recommended for acceptance by X. Yuan. *(Corresponding author: Tianyu Zhang.)*

Tianyu Zhang is with the University of Iowa, Iowa City, IA 52241 USA (e-mail: tianyu-zhang@uiowa.edu).

Jiachen Wang and Song Han are with the University of Connecticut, Storrs, CT 06269 USA (e-mail: song.han@uconn.edu).

X. Sharon Hu is with the University of Notre Dame, Notre Dame, IN 46556 USA (e-mail: shu@nd.edu).

Digital Object Identifier 10.1109/TMC.2025.3599880

to multiple users for data transmissions in each TTI; and ii) choose the MCS index for each user, ensuring that the selected MCS index is identical across all RBs allocated to this user. Therefore, the real-time flow scheduling problem in 5G NR couples together RB allocation and MCS index selection in order to satisfy the timing requirements of industrial flows, making the problem extremely challenging. Time-frequency scheduling for real-time flows in traditional industrial wireless networks has been well studied [9], [10], [11]. In 5G networks, many recent works studied the radio resource allocation problems with objectives to optimize network throughput (e.g., [12], [13], [14]) or age of information (e.g., [15], [16], [17]). Another set of works (e.g., [18], [19], [20]) focuses on scheduling ultra-reliable low-latency communications (URLLC) traffic, especially in the presence of enhanced mobile broadband (eMBB) traffic. However, they mainly consider sparse URLLC traffic, where one can immediately schedule them upon arrival to satisfy the timing requirement by preempting eMBB traffic. In light of improving the real-time performance of 5G networks, [21], [22] 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 [23], [24], [25] study 5G configured grant (CG) scheduling, aiming at providing real-time guarantees for time-critical traffic. However, CG scheduling only applies to 5G UL transmissions and suffers from low flexibility. To the best of our knowledge, this paper is the first work that studies time-frequency DL scheduling in industrial 5G NR, which features per-flow real-time performance guarantee [26]. Specifically, we make the following contributions to this work.

- We leverage a 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., 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 dynamic channel condition changes, 5G-TPS enables online schedule adjustment for affected flows. In large-scale industrial 5G networks with cloud RAN (C-RAN) architecture, 5G-TPS supports user association, respecting the real-time requirements of individual flows.
- Extensive experimental evaluation demonstrates superior performance of 5G-TPS in terms of schedulability ratio compared to the state-of-the-art methods, 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., [27], [28], [29]). However, the MCS index selection in 5G NR scheduling remains an area that lacks comprehensive understanding [30], [31], 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 the 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 and reduced link robustness, 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 5G RAN testbed using OpenAirInterface (OAI) [32], and benchmarked the downlink 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 @3.00 GHz, 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 (see Fig. 1(b)). In the *RFSim* mode, the OAI gNB transmits I/Q samples to the UE over a radio channel simulator, namely the RF simulator, via Ethernet without using the RF boards (i.e., USRP B210 s). In the *over-the-air (OTA)* mode, omnidirectional antennas are connected to the RF boards to transmit signals. In the *SMA cable* mode, two USRP B210 s are directly connected using Sub-Miniature version A (SMA) cables instead of antennas.

We further enrich the throughput measurements in different channel conditions by varying the noise power levels in the RFSim mode and the signal level in the SMA cable mode. Specifically, in the RFSim mode, in addition to the default perfect channel, we enable the OAI `chanmod` option and set the noise power to $-1$ dBm, $-5$ dBm, and $-10$ dBm at the UE side. In the SMA cable mode, we use two 10 dB attenuators and one 30 dB attenuator to configure different reduced amplitude levels (30 dB, 40 dB and 50 dB) of the incoming signal at the UE side by connecting the attenuators in series.

*2) Measurement Settings:* We configure the RAN network to operate on 5G band n78 (3.5 GHz) with 40 MHz of spectrum, which is the maximum bandwidth supported by USRP B210. Since our measurements focus on the 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 for a core network. In the `phy-test` setup, random DL traffic is generated at every scheduling opportunity, and the MAC uses a pre-configured allocation for PDSCH (Physical Downlink Shared Channel) to schedule the DL traffic. We use `iperf` to send UDP traffic from the gNB to the UE in a time duration of 600 seconds, and 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

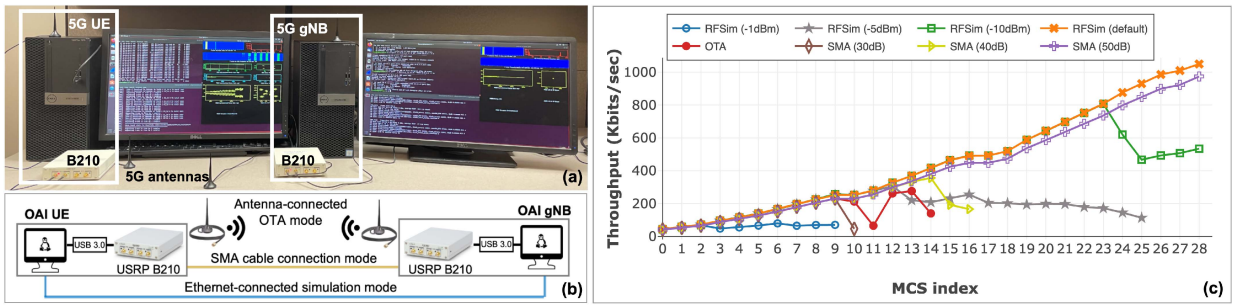



Fig. 1. Motivational experiments on a 5G RAN testbed. (a) Overview of the OAI-based 5G RAN testbed consisting of one gNB and one UE. (b) Architecture of the 5G RAN testbed with three connection modes. (c) Throughput results with varied MCS indices under various connectivity settings.

MCS indices 29, 30, and 31 reserved by 3GPP. Each throughput result is obtained by averaging five independent experimental runs.

### B. Measurement Results

Fig. 1(c) summarizes the throughput results as a function of the MCS index under various connectivity settings. Intuitively, the throughput should increase monotonically with higher MCS indices, as higher modulation schemes and coding rates typically result in higher data rates. However, upon observation, it is apparent that only the results of RFSim with an ideal channel and SMA cable connection with the 50 dB attenuator meet this expectation. All other results show fluctuations as the MCS indices increase. Note that, although the throughput of both setups increases with the MCS index, the throughput under a cabled connection is slightly lower than that of RFSim, indicating an idealized simulation environment of RFSim.

For example, in the RFSim mode, when an extremely high level of noise (-1 dBm) is added to the simulated channel, the network throughput is very low (<80 Kbits/sec). The data also show significant fluctuation when the MCS indices are small (from 0 to 9, i.e., modulation order QPSK). When we further increase the MCS index, the throughput directly drops to 0 (values omitted in Fig. 1(c)) due to an 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. For instance, in RFSim with -1 dBm, -5 dBm, and -10 dBm noise level scenarios, the throughput starts to decrease when the MCS indices are 3, 13, and 24, respectively. Similarly, in the case of cable connection mode with 30 dB and 40 dB attenuators, the throughput decreases when the MCS indices are 10 and 15, respectively.

### 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 the 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 [33], 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.

Second, 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 multi-cell network model and formulate the 5G real-time flow scheduling problem.

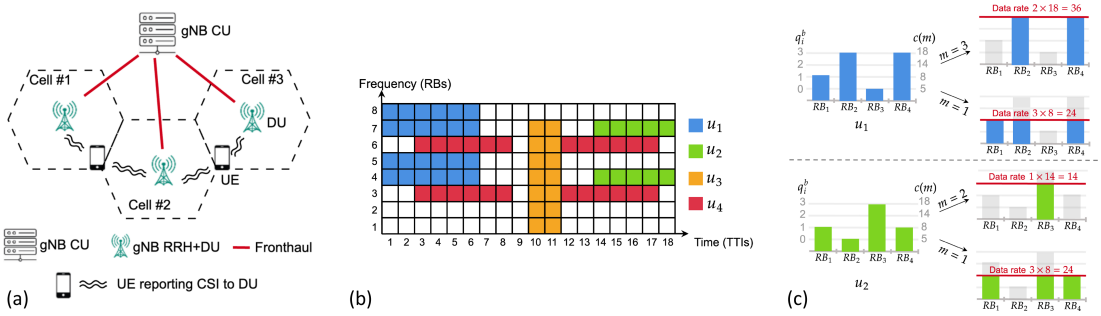



Fig. 2. (a) C-RAN architecture of a multi-cell 5G RAN system. (b) Resource grid in 5G NR. Each block represents a basic time-frequency scheduling unit for UEs. (c) An example of MCS selection for two UEs $u_1$ and $u_2$ on 4 RBs. The colored blocks represent RBs with successfully transmitted data using usable MCS indices. Shaded blocks represent RBs with no data transmission using high MCS indices.

### A. Network and Traffic Model

We consider a multi-cell DL 5G RAN system where one gNB serves a set of $N$ UEs in each cell.

*1) C-RAN:* In large-scale 5G RAN systems across multiple cell sites within a geographical area, a cloud RAN (C-RAN) architecture enables centralized processing and resource management, allowing for more efficient utilization of network resources. Fig. 2(a) gives a C-RAN system comprised of one centralized unit (gNB CU) connecting several distributed units (gNB DUs) with remote radio heads (RRHs). In line with [34], [35], each UE measures the reference symbol received power (RSRP) from the cells that it can hear and creates its CQI measurement set of maximum $Q$ cells it can connect to. The measurement set contains the cell with the highest RSRP, denoted as the primary cell. The other $Q-1$ cells are deemed as secondary cells within the power range of the UE. According to 3GPP [36], the CU performs radio resource management for the entire multi-cell network and determines i) the DU that each UE connects to (i.e., *UE association*) and ii) the resource allocation for all the UEs connected to each DU (i.e., *scheduling*).

*2) OFDMA Resource Grid:* After the UE connects to a certain gNB DU,[1] the communication between the UE and the gNB is through OFDMA-based 5G NR, where network resource is organized as a resource grid that spans both the time and frequency domains as shown in Fig. 2(b). In the time domain, time is equally slotted into transmission time intervals (TTIs), each of which consists of 14 orthogonal frequency-division multiplexing (OFDM) symbols [36]. In the frequency domain, the 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.

*3) 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.[2] 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 units 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.

*4) Resource Allocation Type:* Based on the 3GPP specification [36], 5G NR supports two resource allocation types, *Type 0* and *Type 1*. These types specify how the scheduler allocates RBs for individual flows. In Type-0 resource allocation, a bitmap string is used to specify the RB allocation, where each bit represents the allocation of one RB. In Type-1 resource allocation, RBs are allocated in a consecutive fashion, and it is specified by a start RB index and the total number of RBs allocated. Thus, Type-0 resource allocation is able to provide better flexibility on RB allocation but requires a larger payload size compared to Type-1 resource allocation. According to the 3GPP 5G NR physical layer specification [33], Type-1 resource allocation must be applied when UE receives *Downlink Control Information (DCI) format 1_0* from the gNB, which is used to schedule system information (e.g., Remaining Minimum System Information (RMSI) and Other System Information (OSI)). In other scenarios, the gNB can select resource allocation types according to specific application requirement(s). In this work, we assume that the resource allocation type is determined by the gNB, and we provide scheduling methodologies for both types.

*5) MCS Model:* Besides RB allocation, the scheduler also needs to select a proper MCS for each UE in each TTI [33]. 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 the UE and the 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

[1] In the following, unless otherwise specified, gNB refers to gNB DU.

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

TABLE I
SUMMARY OF IMPORTANT SYMBOLS AND NOTATIONS

| Parameter | Definition | Parameter | Definition |
|---|---|---|---|
| $u_i, f_i, i \in [1, N]$ | UE set and flow set | $s_i^b(t)$ | $s_i^b(t) = 1$ if RB $b$ is allocated to UE $u_i$ in TTI $t$ |
| $\mathcal{B}^+$ | The total set of RBs over the entire bandwidth | $T_i$ | Length of consecutive TTIs of each packet of $f_i$ |
| $P_i, D_i, C_i$ | Period, deadline, and payload size of flow $f_i$ | $R_i(t)$ | The total amount of data transmitted to $u_i$ in $t$ |
| $q_i^b$ | The maximum usable MCS index of flow $f_i$ on RB $b$ | $R_{i,k}$ | The total amount of data transmitted to $u_i$ in packet $p_{i,k}$ |
| $c(m)$ | The achievable data rate under MCS index $m$ | $m_i(t)$ | The selected MCS index for $u_i$ in $t$ |
| $\beta_i(x)$ | The highest data rate function | $S_{j,k}(S_j)$ | TTI duration (set) for packets $p_{j,k}$ (flow $f_j$) in phase 2 |
| $\mathcal{B}_i, \{\mathcal{B}_i^*\}$ | An RB allocation and the candidate set for UE $u_i$ | $\alpha_i(\mathcal{B})$ | The highest achievable data rate with RB allocation $\mathcal{B}$ |

moving between UEs and the gNB [37]. 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 and walls), is non-negligible, and thus, the channel condition between each UE and the gNB varies on different RBs.

We denote $\mathcal{M} = \{0, 2, \ldots, 28\}$ as the set of 29 available MCS indices defined in [33]. 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 data carried on $b$ can be successfully received, and we have $1 \leq q_i^b \leq |\mathcal{M}|$. $q_i^b$ is determined according to the subband CQI submitted by UE $u_i$ on RB $b$, and this CQI value corresponds to a target spectral efficiency based on the used CQI table in [33]. gNB then looks up the MCS index table in [33], where the highest MCS whose spectral efficiency does not exceed the CQI-derived target is selected as $q_i^b$. 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 [33][3]. That is,

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

Note that, although each UE can be allocated multiple RBs in one TTI, it must select and use the same MCS index $m \in \mathcal{M}$ on all the allocated RBs according to the PHY specification of 5G NR [33]. 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. 2(c). 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 index $m = 3$, a higher data rate can be achieved on $RB_2$ and $RB_4$, i.e., $2 \times 18 = 36$. However, a higher MCS index does not always lead to a higher data rate, according to the measurement results in our motivational experiments. For instance, setting a higher MCS index $m = 2$ for UE $u_2$ leads to a lower data rate 14 compared to the data rate 24 that can be achieved by setting $m = 1$.

Thus, 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

$$R_i(t) = \sum_{b \in \mathcal{B}^+} \left( s_i^b(t) \cdot a_i^{b,m_i(t)} \right), \tag{2}$$

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 $R_{i,k} = \sum_{t \in [r_{i,k}, d_{i,k})} R_i(t)$. Table I summarizes the important notations used in this paper.

### B. 5G Real-Time Flow Scheduling Problem

Based on the above system model, the task of the network resource manager at the gNB CU is to 1) perform user association to determine the gNB that each UE connects to and 2) perform flow scheduling to determine a *schedule* for all the UEs connected to each gNB.

*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., $m_i(t)$;

Given the definition of a schedule, the real-time flow scheduling problem in 5G NR is formulated as follows.

*Problem* **P** *(real-time flow scheduling)*: Given the UE set $\mathcal{U}$, flow set $\mathcal{F}$, the modulation and coding rate $c(m)$ on any RB, and the maximum MCS index $q_i^b$ usable by UE $u_i$ on RB $b$, 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 first study the real-time flow scheduling (i.e., Problem **P**) in single-cell 5G RAN networks and outline our 5G NR scheduling framework. Then, we consider a multi-cell C-RAN architecture and propose a UE association approach to be incorporated into our scheduling framework. Below, we first assume that Type-0 resource allocation is applied and the channel condition is stable 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 consider Type-1 resource allocation and extend the scheduling method to account for dynamic channel conditions, where $q_i^b$ for each UE is updated when the channel condition changes.

## IV. PROBLEM ANALYSIS

In this section, we first contrast Problem **P** with traditional real-time scheduling problems to identify its unique challenges and analyze the complexity of Problem **P**. Then, we present an SMT-based exact solution to determine a feasible schedule for Problem **P** and present an overview of the proposed two-phase scheduling framework.

[3] $c(m)$ and $q_i^b$ can be determined through channel estimation using existing methods (e.g., [38]) and are assumed as given to the formulated scheduling problem in this paper.

### A. Comparison With Multiprocessor Scheduling

Problem **P** is markedly distinct from traditional real-time scheduling, particularly the multiprocessor scheduling problems, and their differences are detailed below.

In Problem **P**, each flow $f_i \in \mathcal{F}$ can be considered as a real-time task with period $P_i$, deadline $D_i$ and execution time $C_i$ where executing a task for one time unit on a processor is equivalent to a user transmitting one unit-size data in one TTI. Since flow $f_i$ can be transmitted on an arbitrary number of RBs in one TTI simultaneously, it bears similarities with a preemptable malleable task [39] that can be scheduled for execution on any number of parallel processors. The set of RBs $b \in \mathcal{B}^+$ in the frequency domain can be considered as the set of processors. However, the problems differ in terms of the processing speed models.

On a homogeneous multiprocessor platform, tasks are executed on a set of identical processors, each providing a uniform processing speed $s$. On heterogeneous platforms, the processing speeds of processors are either task-independent [40] or task-dependent [41]. In the former case, each processor $p_b$ has a uniform processing speed $s_b$ for all tasks running on it. While in the latter case, the processing speed of processor $p_b$, denoted as $s_{i,b}$, varies depending on which task (with index $i$) is executed by $p_b$. As a comparison, in Problem **P**, the achievable data rate of flow $f_i$ on each RB $b$ is determined by one additional factor, i.e., the selected MCS level $m$, thus the "processing speed" of RB $b$ can be denoted as $s_{i,b,m}$. Moreover, the fact that each UE must select one MCS level for all the allocated RBs makes Problem **P** more complicated since the "processing speed" of each RB also depends on which RBs are allocated to the flow.

Note that Problem **P** also bears similarities with the DVFS (dynamic voltage and frequency scaling) scheduling problems [42] where the speed of each processor varies by adjusting its voltage and frequency. However, in DVFS scheduling problems, tuning the speed (i.e., voltage/clock frequency) of each processor assigned to a task is typically done independently, whereas in Problem **P**, the MCS level selection for a flow on all its allocated RBs must be identical.

### B. Complexity Analysis

To quantify the complexity of Problem **P**, we first show that Problem **P** under Type-0 and Type-1 resource allocations is both NP-hard. We then analyze its solution space.

Proving that Problem **P** under Type-0 resource allocation is NP-hard can be done by reducing the set packing problem [43] to this problem. Below, we focus on proving the strong NP-hardness of the Type-1 resource allocation case.

*Lemma 1.* Problem **P** under Type-1 resource allocation is NP-hard.

*Proof.* We prove the lemma by reducing 3-partition [44], which is NP-hard in the strong sense, to a special case of this problem. The 3-partition problem is defined as follows: Given a positive integer $B$ and a collection $A = (x_1, x_2, \ldots, x_{3m})$ of positive integers such that $\sum x_i = mB$, $\frac{B}{4} < x_i < \frac{B}{2}$ for each $1 \le i \le 3m$. The 3-partition problem is to determine whether $A$ can be partitioned into $m$ disjoint sets $\{A_1, A_2, \ldots, A_m\}$ such that each $A_k, 1 \le k \le m$ contains exactly 3 elements of $A$ and $\sum_{x_i \in A_k} x_i = B$ for each $1 \le k \le m$.

Given a 3-partition problem, we can construct a special case of the SISO version of Problem **P** under Type-1 resource allocation in polynomial time as follows:

1) The flow set $\mathcal{F} = \{\mathcal{F}_1, \mathcal{F}_2\}$ consists of two sub-flow sets each of which contains $3m$ and $m-1$ flows, respectively. The periods and deadlines of all the flows in $\mathcal{F}$ equal to 1. For each integer $x_i \in A$, we construct a flow $f_i \in \mathcal{F}_1$ with payload size of $x_i, i \in [1, 3m]$. For sub-flow set $\mathcal{F}_2$, we construct $m-1$ flows and each $f_j \in \mathcal{F}_2, j \in [1, m-1]$ has a payload size equal to $C = \max_{x_i \in A} x_i + 1$.

2) There are in total $mB + (m-1)$ RBs in the frequency domain. We use a vector $K = \{\underbrace{1, \ldots, 1}, C, \underbrace{1, \ldots, 1}, C, \ldots, C, \underbrace{1, \ldots, 1}\}$ to denote the amount of data that can be transmitted by all the flows in $\mathcal{F}$ on each RB $b$, where each $\underbrace{1, \ldots, 1}$ represents $B$ consecutive RBs, each of which provides a unit data rate that equals to 1. The number of such consecutive RB sets with unit data rate equals to $m$ and there are $m-1$ RBs each of which provides data rate of $C$.

Next, we show that a feasible RB allocation for flow set $\mathcal{F}$ can be found if and only if there exists a partition of $A$ in the 3-partition problem.

We first prove the "if" direction. If collection $A$ can be partitioned into $m$ disjoint sets $\{A_1, A_2, \ldots, A_m\}$ where each $A_k, 1 \le k \le m$ contains exactly 3 elements of $A$ and $\sum_{x_i \in A_k} x_i = B$ for each $1 \le k \le m$, we can map each flow $f_i \in \mathcal{F}_1$ to an element $x_i \in A$. Then, any 3 flows corresponding to each set $A_k, 1 \le k \le m$ can be allocated with a set of RBs $\underbrace{1, \ldots, 1}$ with totally $B$ unit data rate RBs. Since the data rate requirements of any 3 flows in $A_k$ equals to $B$ and all the RBs in $\underbrace{1, \ldots, 1}$ are with unit data rate, we can allocate a set of consecutive RBs to each flow under Type-1 resource allocation such that the requirements of all the flows are satisfied.

To prove the "only if" direction, we assume flow set $\mathcal{F}$ is schedulable on resource block set $K$. Since $C > \max_{x_i \in A} x_i$, if RB $C$ is allocated to any flow $f_i \in \mathcal{F}_1$, the rest of RBs cannot satisfy the data rate requirements of all the other flows. Because Type-1 resource allocation is applied, each flow must be assigned a set of consecutive RBs. Since the payload size of each flow $f_i \in \mathcal{F}_1$ satisfies $\frac{B}{4} < x_i < \frac{B}{2}$, each set of RBs $\underbrace{1, \ldots, 1}$ must be allocated to 3 flows. Therefore, an RB allocation to flow set $\mathcal{F}_1$ is a feasible partition of $A$. This completes the proof. □

Next, we analyze the solution space of Problem **P**. The gNB needs to allocate $B$ RBs among $N$ UEs and assign each UE an optimal MCS (among 29 possible levels) in each TTI within the hyper-period $H$. This gives a total number of $(N^B \cdot (29)^N)^H$ possibilities in the solution space. Consider a typical industrial 5G NR system, this number can be on the order of $10^{4.8e4}$ (e.g., $N = 50$, $B = 100$, $H = 200$).

### C. SMT Formulation

Given the NP-hardness of Problem **P** which exhibits combinatorial characteristics, we present an approach based on the

Satisfiability Modulo Theory (SMT) to determine a feasible schedule for the flow set. According to the definition of schedule in Definition 1, we define RB allocation $s_i^b(t)$ and MCS level selection $m_i(t)$ for each flow $f_i$ as integer variables in the SMT context and generate assertions that correspond to the scheduling constraints described in Section III as follows.

*Data amount constraint*. Given the timing requirement of each flow $f_i \in \mathcal{F}$, i.e., at least $C_i$ amount of data must be transmitted within $[r_{i,k}, d_{i,k})$ for each released packet $p_{i,k}$, we have

$$\forall i, \forall k : R_{i,k} \geq C_i.$$

*RB allocation constraint*. If Type-0 resource allocation is applied, no additional constraint is needed. Otherwise, under Type-1 resource allocation, each flow must be assigned a set of consecutive RBs in each TTI. Thus, we have

$$\forall i, \forall t, \forall b', b'' \in [1, N-1], b' \neq b'' :$$
$$\neg((s_i^{b'+1}(t) - s_i^{b'}(t) = 1) \wedge (s_i^{b''+1}(t) - s_i^{b''}(t) = 1)).$$

Based on the above formulation, an SMT solver (e.g., Z3) can be applied to find a feasible schedule (if exists) to satisfy the timing requirements of all the flows. However, the SMT-based solution suffers from severe scaling challenge due to the extremely large search space of Problem **P**. The running time of the SMT-based approach explodes quickly when the number of flows and RBs becomes large. Experimental results in Section VIII demonstrate this point quantitatively. To tackle this, the main focus of this work is to design an efficient and effective scheduling framework that can be used in large-scale 5G RAN systems.

### D. Overall Scheduling Framework

In this work, 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 MCS index. This incurs large control resource overhead, which, in turn, reduces the amount of network resources allocated to PDSCH for transmitting actual data. 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). Thus, 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. Section V describes the details of Phase 1 of 5G-TPS. 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. Section VI describes the details of Phase 2 of 5G-TPS.

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-B.

## 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?

### A. Flow Set Schedulability

To answer the first question, we introduce a lemma to help define the feasible RB allocation for each flow (i.e., answering the first question in Phase 1 design).

*Lemma 2.* If the amount of transmitted data per TTI is larger than or equal to $\left\lceil \frac{C_i}{D_i} \right\rceil$, flow $f_i$ is schedulable, i.e., satisfies the deadline.

The proof of Lemma 2 is straightforward and thus omitted. According to Lemma 2, 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 $\left\lceil \frac{C_i}{D_i} \right\rceil$. Based on Lemma 2, 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 2.

*Constraint 2.* Each RB $b \in \mathcal{B}^+$ can be allocated to at most one flow in each TTI.

According to Lemma 2, Constraint 1 guarantees that each flow $f_i$ is schedulable with allocated RBs in $\mathcal{B}_i$. Thus, the theorem apparently holds. Constraint 2 guarantees that no transmission interference occurs between any two UEs. To find a feasible RB allocation $\Theta$ 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 $\{\mathcal{B}_i^*\}$. Below, we elaborate on 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., $\sum_{x=1}^{B} \binom{B}{x} = 2^B - 1$

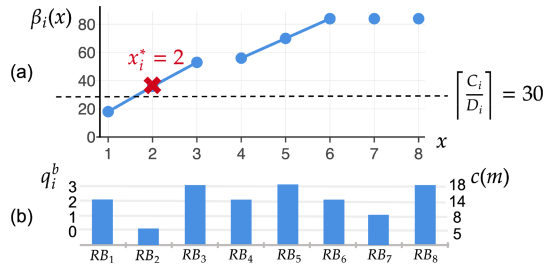



Fig. 3. (a) An example of the highest data rate function $\beta_i(x)$. The dashed line represents the required data rate. (b) The channel conditions of user $u_i$ on a network with 8 RBs.

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 flow $f_i$, and only include in it the most promising RB allocations, instead of solving in one shot the RB allocation 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 2, the feasible RB allocations must have achievable data rate larger than or equal to $\left\lceil \frac{C_i}{D_i} \right\rceil$. 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 fewer 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 in identifying all 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 $f_i$ 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. 3 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 equal to 2 where the achievable data rate is $4 \times 14 = 56$. Thus, we have $\beta_i(4) = 56$.

As depicted in Fig. 3, all the RB allocations resulting $\beta_i(x) \geq \left\lceil \frac{C_i}{D_i} \right\rceil$ $(x \leq B)$ 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. 3, $\beta_i(2) = 36$ 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 3.* 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 $\zeta_j$. 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 index $\zeta_j$ is greater than the current value $x$, $\zeta_j$ 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 $\zeta_j$, a smaller MCS index $\zeta_h$ has to be selected and the slope of $\beta_i(x)$ reduces to $c(\zeta_h)$. □

*Lemma 3* 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. 3, $\beta_i(4) = \alpha_i(\{1,3,4,5\})$ and the MCS index used is $m = 2$, i.e., $c(2) \cdot 4 = 14 \times 4 = 56$. 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., $\beta_i(x) \geq \left\lceil \frac{C_i}{D_i} \right\rceil$. For instance, in Fig. 3, 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 4.* 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 3, $\beta_i(x') = c(q_i^{b_j}) \cdot x'$ 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 4* 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. 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. 3, the set of considered number of allocated RBs is further reduced from $x \in \{2, 3\}$ to $x = 2$. Based on Lemma 3&4, 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 $\beta_i(x) \geq \left\lceil \frac{C_i}{D_i} \right\rceil$, 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. 3, $x_i^* = 2$ and $\{\mathcal{B}_i^*\} = \{\mathcal{B} | \alpha_i(\mathcal{B}) = \beta_i(2)\} = \{\{3,5\},\{5,8\},\{3,8\}\}$.

### 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 Constraint 2 is 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) Constraint 2 is 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]). The high-level idea is to determine the RB allocation for flows in the increasing order of their candidate set size (i.e., $x_i^*$). In each iteration for $f_i$, we select the first RB allocation $\mathcal{B}_i \in \{\mathcal{B}_i^*\}$ satisfying Constraint 2 given all the RB allocations of the previously scheduled flows. The time complexity is $O(NWB)$ where $W$ is the maximum size of $\{\mathcal{B}_i^*\}$ among all the flows.

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

In Phase 1, each flow $f_i$ is allocated the same set of RBs in all the TTIs within the hyperperiod. This may allocate unnecessary RBs for certain flows in the time domain, i.e., some redundant RBs may be allocated in certain TTIs following the same RB allocation setting in Phase 1. 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 real-time requirements of the unschedulable flows by using those RBs 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.* 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 2 (i.e., $\beta_i(x^*) > \left\lceil \frac{C_i}{D_i} \right\rceil$), $f_i$ only needs $\left\lceil \frac{C_i}{\beta_i(x_i^*)} \right\rceil$ TTIs to complete the transmission of each released packet, where the RBs in the rest $P_i - \left\lceil \frac{C_i}{\beta_i(x_i^*)} \right\rceil$ TTIs are not used. As an example, the white blocks in Fig. 4 represent the remaining RBs in the time domain.

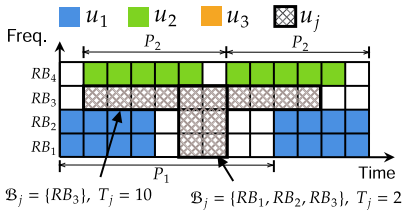



Fig. 4. Illustration of Phase 2 scheduling for flow $f_j$ using the remaining RBs. The colored blocks represent the RB allocation for scheduled flows in Phase 1. The white blocks represent the remaining RBs. The patterned blocks represent two feasible RB allocations for $f_j$.

### B. Phase 2 Overview

In Phase 2, we use the remaining RBs to generate a feasible schedule for each unschedulable flow $f_j \in \mathcal{F}_2 = \mathcal{F} - \mathcal{F}_1$, 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 and the scheduled TTIs. In the time domain, we schedule each packet of flow $f_j$ 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 by specifying the start TTI index and the number of TTIs according to 3GPP specification [46].

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 $\mathcal{B}_j$ (with the optimal MCS index $m_j$) in the frequency domain, and TTI duration (denoted as $S_{j,k} = [t_{j,k}, t_{j,k} + T_j)$) in the time domain, 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 flow $f_j$ 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 $\mathcal{F}_2$.

*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 3.* 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 $\alpha(\mathcal{B}_j) \cdot T_j \geq C_i$.

*Constraint 4.* $\mathcal{B}_j$ cannot share a common RB in a TTI with any $\mathcal{B}_i$ ($f_i \in \mathcal{F}$).

*Proof Sketch.* Constraint 3 guarantees that the total amount of data transmitted over RBs $b \in \mathcal{B}_j$ in $[t_{j,k}, t_{j,k} + T_j)$ is larger than or equal to $C_i$. Constraint 4 guarantees the interference-free transmissions of UEs.

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

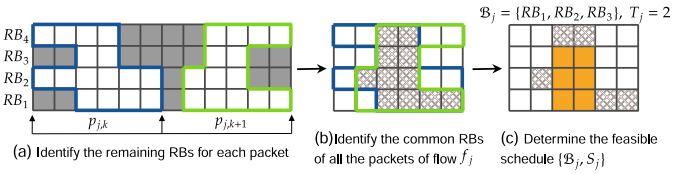



Fig. 5. Illustration of the feasible schedule generation for flow $f_j$. (a) The white blocks represent the remaining RBs of packet $p_{j,k}$ and $p_{j,k+1}$. (b) The patterned blocks represent the common RBs of the two packets. (c) The orange blocks represent the determined feasible schedule for $f_j$.

(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 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 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[4] 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 meet the deadline of each $p_{j,k}$, i.e., satisfying Constraint 3.

To satisfy requirement (i), we traverse the set of remaining RBs usable by the packets $p_{j,k|k=1,2,\ldots}$ and identify the common RBs (i.e., RBs in the same relative TTIs in each period) of all the packets (see Fig. 5(a) and (b)). For requirement (ii), generating a feasible schedule satisfying Constraint 3 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. 5(c)). Here, 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].

The main functions in Phase 2 are summarized in Algorithm 1. The computational cost of the algorithm is dominated by line 6, where the largest empty rectangle search is invoked once per packet. Given the identified common RBs in a resource grid of at most $H$ TTIs and $B$ RBs, the search runs in $O(HB\log(HB))$ time according to [47]. With at most $|\mathcal{F}_2| \cdot \frac{H}{P_{min}}$ packets handled in the hyperperiod, the time complexity of Algorithm 1 is $O(|\mathcal{F}_2| \cdot \frac{H}{P_{min}} \cdot H \cdot B\log(HB))$.

As proved in Section IV-B, Problem **P** is NP-hard in the strong sense, so 5G-TPS aims to generate a schedule for flow set $\mathcal{F}$ in an efficient and effective manner. For this reason, some feasible solutions of Problem **P** may be pruned from the search space. For example, Phase 1 accepts only the smallest RB set that meets

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

**Algorithm 1:** Schedule Generation in Phase 2.

1: Sort flows in $\mathcal{F}_2$ in the decreasing order of flows' utilization;
2: Determine the set of remaining RBs within $[1, H]$;
3: **for** $f_j \in \mathcal{F}_2$ **do**
4: **for** each packet $p_{j,k}$ **do**
5: Identify the common RBs of all the packets of $f_j$;
6: Determine the feasible schedule $\{\mathcal{B}_j, S_j\}$ for flow $f_j$;
7: **end for**
8: Update the set of remaining RBs;
9: **end for**

a flow's data-rate demand, and Phase 2 sequentially generates schedules for flows in decreasing-utilization order. Therefore, our solution only provides a sufficient schedulability condition for flow set $\mathcal{F}$. That is, a schedulable flow set $\mathcal{F}$ may be deemed as unschedulable by 5G-TPS, i.e., Phase 2 fails to find a feasible schedule for any flow $f_j \in \mathcal{F}_2$.

## VII. Extending the Solution Framework

In the previous sections, we consider a single-cell 5G RAN under Type-0 resource allocation, assuming that the channel conditions of individual UEs are stable. In this section, we generalize our system model and extend the 5G-TPS framework by considering multi-cell 5G RAN under Type-1 resource allocation with dynamic channel conditions.

### A. Type-1 Resource Allocation

In the following, we describe how to solve Problem **P** under Type-1 resource allocation (i.e., each UE is allocated with a set of consecutive RBs within each TTI) using 5G-TPS.

The constraint posted by Type-1 resource allocation introduces a key difference from Type-0 for the RB allocation of each packet in the frequency domain, i.e., the calculation of the highest data rate function $\beta_i(x)$. Since the RB allocation under Type 1 is a special case of that under Type 0, the properties regarding $\beta_i(x)$ (i.e., Lemmas 3&4) still hold. Therefore, below we focus on describing the differences in the calculation of $\beta_i(x)$ such that a feasible schedule under Type-1 resource allocation can be found according to Theorem 1.

According to the description in Section V, $\beta_i(x)$ is calculated by selecting the maximum $x$ RBs $\mathcal{B} \subseteq \mathcal{B}^+$ and $|\mathcal{B}| = x$. Then, $\beta_i(x) = \alpha_i(\mathcal{B})$. Under Type-1 resource allocation, the only difference is to select the RB allocation $\mathcal{B}$ containing $x$ consecutive RBs. This can be done by enumerating at most $B$ possible RB allocations, each of which starts from an RB $b \in \mathcal{B}^+$. Then, the candidate set containing multiple RB allocations with consecutive RBs is constructed in phase 1. Another noteworthy difference exists in phase 2 when we generate a feasible schedule using the identified common RBs. The requirement of consecutive RBs constrains the target to be a rectangle instead of a logical one as in Phase 1. Type-1 resource allocation imposes an additional RB allocation constraint to Problem **P**. Thus, the schedulability of a flow set under Type 1 is inevitably lower than that under

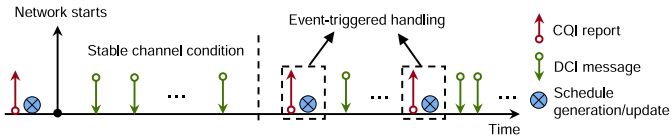



Fig. 6. The network execution model for handling dynamic channels.

Type-0 resource allocation, also to be shown in the evaluation section. However, as discussed in Section III, Type-1 resource allocation is mandatory in certain scenarios (e.g., system information transmissions). Our two-phase scheduling framework supports both types of resource allocation and is more practical for deployment in real industrial 5G applications.

### B. 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. 6, when the network channel condition is stable, the schedule of each UE is carried out through the DCI messages on PDCCH (Physical Downlink Control Channel) to meet the timing requirement of each flow. 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.

The gNB may receive multiple updated CQI reports from different UEs within a short time interval before completing the updated schedule generation. 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 $f_i \in \mathcal{F}$, (i.e., the RB allocation $\mathcal{B}_i$ and the TTI assignment $\{[t_{i,k}, t_{i,k} + T_i | k = 1, 2, \ldots\}$), determine the schedule adjustment to meet the deadlines of all the flows $f_i \in \mathcal{F}$.

The channel condition change of $u_e$ consists of three cases.

*Case 1:* $\forall b \in \mathcal{B}_e, a_e^{b,m} = c(m)$. In this case, the maximum usable MCS index $q_e^b$ 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 satisfy the data rate requirement according to Lemma 2. 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 $\exists m', \alpha(\mathcal{B}_e) \geq \left\lceil \frac{C_e}{T_e} \right\rceil$. In this case, the maximum usable MCS index $q_e^b$ on certain allocated RB(s) $b$ is smaller than the selected 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., $\alpha(\mathcal{B}_e) \geq \left\lceil \frac{C_e}{T_e} \right\rceil$. Thus, the gNB only updates the MCS index in the updated schedule for $u_e$.

*Case 3:* $\forall m \in [0, 28], \alpha(\mathcal{B}_e) < \left\lceil \frac{C_e}{T_e} \right\rceil$. 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 use the remaining RBs to adjust the RB allocation of $f_e$ to meet its timing requirement. Specifically, we identify all the remaining RBs along with the RBs allocated to $f_e$ and perform RB re-allocation for $f_e$ 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.[5] If Phase 1 and Phase 2 fail, we deem that flow $f_e$ is unschedulable after channel condition change.

*Schedule Update Frequency.* In harsh industrial environments, channel conditions may fluctuate frequently, making frequent schedule regeneration costly due to both computational overhead (recomputing schedules) and communication overhead (disseminating updates via DCI messages). As a trade-off, our framework can incorporate a conservative channel state estimation strategy by using a lower-bound (i.e., more conservative) estimate of the maximum usable MCS index. This approach increases the tolerance of the generated static schedules to moderate channel variations and reduces the need for frequent rescheduling.

**Algorithm 2:** User Association.

```
Sort UEs in the decreasing order of flows' utilization;
for u_i ∈ U do
    Sort G_i in the decreasing order of channel conditions;
    for g_ik ∈ G_i do
        if a feasible schedule can be generated for u_i and all
          the UEs associated to g_ik by 5G-TPS then
            u_i is associated to g_ik;
        end if
    end for
end for
```

### C. C-RAN User Association

In a large-scale 5G RAN system with the C-RAN architecture, the gNB CU needs to perform user association to determine the gNB DU that each UE connects to. Ideally, all UEs would connect to their primary cells with the best channel conditions. However, in practice, this could lead to overload in certain cells, resulting in failure to meet the timing requirements of specific flows. Many studies in the literature have explored resource allocation for 5G C-RAN (e.g., [48], [49], [50]), including aspects of user association. However, these studies primarily focus on enhancing network throughput, increasing data rates,

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

and optimizing energy consumption rather than addressing the stringent real-time requirements of industrial applications. Our objective differs as we seek a user association method that can effectively integrate with our proposed 5G-TPS framework while satisfying the real-time requirements of all the flows. The UE association problem can be defined as follows.

*Problem P4*. Given the UE set $\mathcal{U}$, the cell set $\mathcal{G}_i = \{g_{i1}, g_{i2}, \ldots, g_{iQ}\}$ that each UE $u_i$ can connect to based on the CQI measurement information, the flow set $\mathcal{F}$ and all the input specified in Problem 1, determine the gNB DU that each UE connects to such that the real-time requirements of all the flows are satisfied.

The solution space of Problem P4 can be extremely large since its optimization version (i.e., maximizing the flows satisfying their deadline requirements) has non-convex and combinatorial structure. Therefore, we propose an efficient and effective heuristic integrated with 5G-TPS to determine the user association and schedules for all the UEs connecting to each gNB. If we simplify all the gNBs as bins of unit size and each UE with the corresponding flow as an item of size between 0 and 1, Problem P4 becomes the Bin Packing problem. Thus, we can leverage the insight of first-fit-decreasing (FFD) algorithm [51] which is an effective approximation solution.

The high-level idea of the proposed user association heuristic is to prioritize associating each UE to the gNB with the best channel condition, as long as the resources of this gNB can satisfy the real-time requirements of all the flows communicating with it. Algorithm 2 shows the pseudo-code of the heuristic. We prioritize UEs based on the utilization of their corresponding flows (i.e., $C_i/P_i$) since flows with higher utilization are generally more challenging to schedule. At each iteration for UE $u_i$, we traverse the gNBs in $\mathcal{G}_i$, ordered from best to worst channel conditions. If 5G-TPS can generate a feasible schedule for the $u_i$ and all other UEs already associated to a certain gNB, $u_i$ is then associated to that gNB.

## VIII. Performance Evaluation

This section presents the experimental evaluation through extensive simulations to evaluate the proposed 5G-TPS framework. Although we built a real-world 5G RAN testbed, as described in Section II, conducting 5G-TPS performance evaluations on the testbed is not feasible for two reasons. From the hardware aspect, the current 5G testbed consists only of one gNB and one UE, and the high cost of USRP devices makes it difficult to create a large-scale testbed for the 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 measures the channel condition and reports a 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[6] and out of the scope of this work.

### A. 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. We consider a 10 MHz bandwidth network consisting of 50 RBs, i.e., $B = 50$. Each real-time traffic flow $f_i$ is randomly generated with payload size $C_i$ and period (deadline) $P_i(D_i)$ drawn from the uniform distributions over $[20, 1024]$ bytes and $[1, 20]$ ms, respectively.

*1) Variables:* The variables used in the experiments include the number of RBs $B$, the number of flows $N$ and the normalized flow set utilization $U^* \in (0, 1]$ where $U^* = \sum_{f_i \in \mathcal{F}} \frac{C_i}{P_i(B \cdot c(|\mathcal{M}|))}$. Here, $U^*$ captures the flow set workload on one resource block with the maximum modulation and coding rate $c(|\mathcal{M}|)$. $U^* = 1$ means that the flow set is potentially schedulable only if the maximum MCS level $|\mathcal{M}|$ can be used by each UE on all RBs $b \in \mathcal{B}^+$ (i.e., under ideal channel conditions). Type-0 and Type-1 resource allocation settings are evaluated separately.

*2) Metrics:* We use the following evaluation metrics. First, in the stable channel condition, we use the *Schedulability Ratio (SR)* to evaluate the performance of 5G-TPS in finding a feasible schedule. SR is defined as the ratio of feasible flow sets according to Theorem 1 to all the generated flow sets; In the dynamic channel condition, we use the *number of Deadline Missed flows (DM)* to evaluate the effectiveness of 5G-TPS in schedule adjustment. In addition, we evaluate the channel efficiency of all the methods by comparing the average Transport Block size (TBS) and the number of used RBs by individual methods across the entire hyperperiod. Since the focus of this work is studying the traffic real-time performance, another metric used is the average latency of all the flows.

*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 [52]. 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 network slicing under fixed-priority scheduling [21]. CA is based on an over-simplified resource model where the entire network bandwidth is treated as one single RB. A generalized version considering multiple RBs, denoted as *CA-Ext*, is also implemented for the evaluation comparison.[7]

**RR**, **MT**, and **PF**: Three built-in flow schedulers (i.e., round-robin, maximum CQI, and proportional fair) in OAI's 5G RAN implementation [53].

**DRR** and **PQ**: Two extended flow schedulers (i.e., deficit round-robin and priority queue) based on the built-in scheduling algorithms [54].

[6]https://lists.eurecom.fr/sympa/arc/openair5g-user/2023-01/msg00105.html.

[7]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 achieved by CA-Ext.

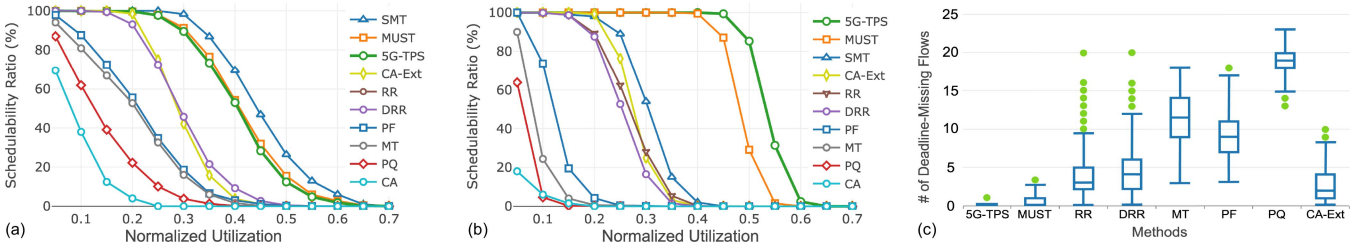



Fig. 7. Evaluation results under Type-0 resource allocation. (a) SR comparisons in small networks, where each point of an individual method represents the average value of 5000 trials. (b) SR comparisons in large networks. (c) DM comparisons in large networks, where the box plots depict the DM distributions of all the methods across 200 hyperperiods and the green dots represent the outliers.

### B. Performance Results

*1) Stable Channel Condition:* In the first set of experiments, we compare the SRs of all the methods by varying the normalized flow set 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 and 8 RBs, and (2) a more practical large-scale network with $N = 25$, and $B = 50$. Further, we set a $2700s$ timeout limit for the SMT approach in the large-scale network setting to prevent it from spending a long time finding a result.

*Small scale networks*. Fig. 7(a) shows the SR as a function of utilization $U^*$ 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 increase in $U^*$ (the curves of RR and DRR overlap due to their same SR results), and SMT dominates others as an exact solution. The SR gap between SMT and 5G-TPS is very 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 (0.79% lower on average) with MUST. However, the performance of MUST drops significantly when the network scales to the normal size (i.e., $N = 25$, $B = 50$) to be shown in the next set of experiments. Note that 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 [21], where the entire network bandwidth is treated as one single RB.

*Large scale networks*. Fig 7(b) shows the SR as a function of $U^*$ in large scale networks and all the curves remain similar trends to those observed in the small scale networks. In the large-scale network setting, the number of UEs increases, which increases the system workload. On the other hand, the increase in RBs and antennas provides more network resources that can benefit the SR. From the results, we can observe 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 SR drop of SMT is mainly because it fails in most cases under the timeout limit due to the extremely large search space. 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 if more network resources are available.

On the other hand, the SRs of both 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 satisfying their deadlines.

*2) Dynamic Channel Condition:* In the second set of experiments, we evaluate the performance of all the methods in large-scale networks (i.e., $N = 25$ and $B = 50$) with dynamic channel conditions. We randomly generate one flow set with normalized utilization $U^* = 0.4$ and run continuously for 200 hyperperiods. The maximum usable MCS index $q_i^b$ 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 online dynamic schedule adjustment and the performance of CA is dominated by CA-Ext according to the results in the previous experiments.

Fig. 7(c) shows the DM distributions of all the methods based on box plots and the result of each method represents the statistics of DM in one hyperperiod where 25 flows run in the network. We can observe from the results that all the methods suffer from deadline misses for certain flows in dynamic channel conditions. However, 5G-TPS outperforms all the other approaches 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 generates flows missing deadlines in 66 hyperperiods out of 200 hyperperiods. The other methods suffer from higher DM, especially for MT, PF, and PQ, where flows miss their deadlines in all the hyperperiods.

*3) Channel Efficiency:* Channel efficiency is a crucial performance metric in 5G scheduling and resource management. Since any scheduling method typically has a specific optimization goal (e.g., throughput, latency, or fairness), achieving higher channel efficiency, i.e., making more effective use of limited network resources, makes it easier to meet the intended optimization objectives. In this experiment, we evaluate channel efficiency across all methods using two key metrics: average TBS and the total number of RBs used within a hyperperiod. The parameter

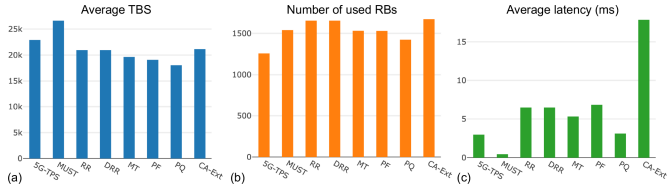



Fig. 8. Channel efficiency comparisons under Type-0 resource allocation.

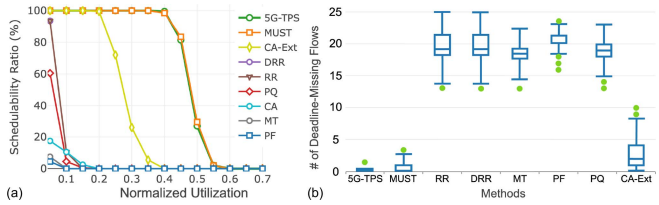



Fig. 9. Evaluation results under Type-1 resource allocation. (a) SR comparisons (SMT is excluded due to its extremely large runtime overhead). (b) DM comparison results are shown by the box plots with outliers, i.e., the number of deadline-missing flows in extreme cases.

settings follow those of the previous large-scale network experiments, with flow set utilization fixed at $U^* = 0.4$. For average TBS, we determine the MCS using Table 5.1.3.1-3 in [33] and then compute the TBS for all UEs in each TTI before averaging the results. For RB usage, since the hyperperiod is 40 and there are 50 RBs per TTI, the total number of available RBs in the hyperperiod is 2000.

Fig. 8(a) presents the experimental results as a bar chart, where the number of RBs used is enlarged by 10 times to improve visibility. From the figure, it is evident that MUST achieves the highest average TBS. This is expected, as MUST prioritizes allocating RBs with the highest data rate when assigning resources to UEs. 5G-TPS achieves a slightly lower average TBS than MUST (it still outperforms other methods). However, when considering the schedulability results in Fig. 7(b), it becomes clear that simply maximizing channel efficiency does not necessarily lead to better real-time performance, meaning it does not always satisfy a greater number of flows' timing requirements.

Regarding the number of RBs used, the results in Fig. 8(b) show that 5G-TPS consumes the fewest RBs within the hyperperiod while achieving the highest schedulability. This further confirms that 5G-TPS effectively utilizes network resources to meet real-time requirements. From another perspective, all methods exhibit relatively low RB utilization. For instance, CA-Ext achieves only 83.6% RB utilization, yet its schedulability drops to zero at $U^* = 0.4$. This underscores the challenge of meeting real-time requirements, highlighting the need for efficient scheduling strategies to properly allocate resources.

*4) Average Latency:* In this set of experiments, we evaluate the timing performance of all the methods by comparing the average latency of all the flows in the hyperperiod. We still follow the large-scale network setting, with utilization $U^* = 0.4$. If a packet scheduled by a method misses its deadline, it is dropped. In this case, the latency is calculated to be the deadline of the flow.

The average latency results are shown in Fig. 8(c), and similarly, the numbers are enlarged by 1000 times for improved visibility. We can see that MUST achieves the lowest average latency, followed by 5G-TPS, while CA-Ext has the highest average latency, primarily due to its over-simplified single-RB model. When comparing the average latency results with the schedulability results from Fig. 7(b), it becomes evident that lower latency does not always correlate with higher schedulability. For example, 5G-TPS has a higher average latency than MUST, yet its schedulability is better. Similarly, while PQ achieves relatively low average latency, its schedulability performance is poor.

These observations lead to an important conclusion: reducing latency does not necessarily improve schedulability. We will further elaborate on this distinction in Section IX, where we provide a more detailed analysis of the relationship between latency reduction and real-time performance.

*5) 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. We compare the average runtime of all the methods with the number of UEs (i.e., $N$) and RBs (i.e., $B$) as variables. From the results, we observe a linear increase in the runtime of most methods with the increase of $N$ and $B$. However, the runtime of MUST experiences 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.

*6) Type-1 Resource Allocation:* We perform SR and DM comparisons for large-scale networks under Type-1 resource allocation, and the results are shown in Fig. 9. The additional resource allocation constraint, i.e., each UE is allocated with a set of consecutive RBs within each TTI, reduces the flexibility of resource allocation, which naturally leads to a decrease in schedulability. Notably, in Fig. 9(a), the SR of MUST and CA-Ext show almost no change compared to the results in Fig. 7(b). This is because these two methods inherently allocate RBs to UEs sequentially in the frequency domain, already satisfying the Type-1 constraint. In contrast, the SR of all other methods experiences varying degrees of decline due to the added restriction (the curves of RR and DRR overlap due to their same SR results). However, our proposed 5G-TPS demonstrates a much smaller average decline in SR (9.89%) compared to other methods (76.7%). On the other hand, while the SR of 5G-TPS is almost identical to that of MUST under the Type-1 setting, Fig. 9(b) shows that under dynamic channel conditions, 5G-TPS still outperforms all other methods, including MUST, in terms of achieving lower DM, with much lower runtime overhead as stated in Section VIII-B5. In contrast, the DM of other methods is further significantly reduced due to the additional constraints imposed by Type-1 resource allocation.

*7) C-RAN:* In this set of experiments, we evaluate 5G-TPS in multi-cell networks by incorporating the user association algorithm in Algorithm 2. User association heuristic design in multi-cell networks is typically tied to the specific optimization objectives and takes the channel condition across different cells

TABLE II
SR COMPARISON RESULTS UNDER DIFFERENT SETTINGS OF $N$ AND $U^*$

| $(N, U^*)$ | (10, 0.8) | (10, 1.2) | (10, 1.6) | (25, 1.2) | (25, 1.6) |
|---|---|---|---|---|---|
| 5G-TPS | 100 | 96.5 | 49.5 | 99.5 | 78.5 |
| MUST | 95.5 | 40.0 | 2.5 | 75.5 | 9.5 |

into consideration. Given the lack of user association methods specifically considering timing performance guarantees, as described in Section VII-C, we do not directly compare 5G-TPS to a resource allocation method with diverse optimization objectives. Instead, for the sake of fair comparisons, we extend MUST, the most comparable method based on the previous experiment results, by incorporating the user association heuristic in [50]. Specifically, for each UE, we associate it to its primary cell and check whether all the UEs are schedulable by MUST. If not, this process repeats by associating the UE to the secondary cell with the best channel condition.

We apply Type-0 resource allocation and set the number of gNB to 3. We randomly generate the location of each UE and accordingly determine the CQI measurement set of the three cells. We vary the number of UEs, $N$, and the normalized flow set utilization $U^*$. Note that since we have multiple cells, each of which can serve a set of UEs, we allow the flow set utilization to be larger than 1.

Table II summarizes the SR comparison results under different settings of $N$ and $U^*$. We have three observations. 1) At the highest level, the performance of 5G-TPS, in terms of schedulability, dominates that of MUST. For example, when $N = 25, U^* = 1.6$, the SR of 5G-TPS (78.5%) is 726% higher than that of MUST (9.5%). 2) When comparing the results with those in a single-cell setting in Fig. 7(b), we observe that the performance of both methods declines in the multi-cell scenario. This is mainly caused by suboptimal user association. For instance, in the single-cell setting with $U^* = 0.5$, the SR of 5G-TPS and MUST are 84.3% and 29.6%, respectively. However, in the multi-cell setting, roughly corresponding to $U^* = 1.6$, the SR drops to 78.5% for 5G-TPS and 9.5% for MUST. Notably, the decline for 5G-TPS is much smaller than for MUST, indicating that the user association strategy in Algorithm 2 is more effective than the greedy method, which relies solely on channel quality for user association. 3) The results in Table II show that as $N$ increases, the SR also improves. For example, the SR of 5G-TPS increases from 49.5% when $N = 10$ to 78.5% when $N = 25$. This improvement is primarily because, at the same $U^*$, a higher number of UEs results in a smaller workload per UE. This increases the flexibility of user association, making it easier to find feasible associations that meet the timing requirements.

## IX. REAL-TIME PERFORMANCE OF 5G IIOT

There has been extensive research in the literature on 5G timing performance, particularly after 3GPP introduced the concept of URLLC [55]. The primary goal of these studies has been to reduce packet latency. However, none of these works can provide real-time guarantees for 5G RAN. The fundamental reason lies in the misconception of real-time performance in the literature: the assumption that real-time requirement is equivalent to low-latency requirement (put differently, the lower the latency, the better the real-time performance). For example, a recent study [56] measuring industrial 5 G's real-time performance defines hard and soft real-time requirements according to the latency thresholds: traffic with latency requirement below 100 ms is classified as hard real-time, while traffic with latency requirement above 100 ms is considered soft real-time. In reality, real-time performance is not determined by the absolute latency value but rather by whether every instance of every flow meets its deadline throughout the system operation [57]. Hard real-time and soft real-time are thus differentiated by whether the system strictly prohibits any packet from missing its deadline or allows occasional deadline misses.

Below, we summarize existing literature on 5G timing-related research and explain why these approaches fail to provide real-time performance guarantees.

One category of 5G timing-related research focuses on reducing latency at the PHY layer, introducing solutions such as new or modified frame structures, waveform designs, or improved modulation schemes [58], [59], [60], [61], [62]. These approaches theoretically reduce latency in the best-case scenario, e.g., adopting shorter TTI to reduce achievable latency.

Another major portion of research focuses on latency reduction at the MAC layer, primarily through user scheduling and resource allocation [20], [63]. 5G scheduling can be broadly categorized into channel-independent scheduling and channel-dependent (dynamic) scheduling. The former allocates radio resources equally across all users, while the latter optimally assigns resources based on users' channel conditions, typically with the goal of minimizing latency. However, these optimization-based latency reduction approaches can only evaluate average latency through simulation experiments or analyze latency distributions over a large number of experimental runs. The results are highly dependent on experimental settings and provide no formal guarantee on worst-case latency. Some work attempts to address the real-time requirement by proposing deadline-aware scheduling algorithms that allocate resources to the most urgent packets [52], [64], [65]. Such methods, however, are essentially best-effort approaches, meaning they cannot determine whether real-time requirements are met, let alone provide any theoretical guarantees.

Another category of research focuses on URLLC scheduling [18], [19], [20], particularly on co-scheduling URLLC and eMBB traffic. These works typically assume that sufficient network resources are available for URLLC, and their primary goal is to maximize eMBB throughput while ensuring URLLC timing requirements are met under this assumption.

Existing 5G timing performance research is insufficient for providing real-time guarantees because it typically optimizes latency without strict timing determinism and relies on average-case latency statistics. From the experimental results, we can also observe that reducing latency does not inherently satisfy real-time performance requirements. Real-time performance is not just about minimizing latency but about properly accommodating network resources while ensuring that each individual packet meets its deadline. Without an explicit

mechanism to respect individual packet deadlines, even a system with low average latency may still experience deadline misses, rendering it unsuitable for mission-critical industrial applications.

## X. Conclusion and Future Work

In this paper, we leverage a 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 guarantee through time-frequency 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 changing. In large-scale industrial 5G networks with C-RAN architecture, 5G-TPS supports user association, respecting the real-time requirements of individual flows. 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.

## References


[1] 3GPP, "Service requirements for cyber-physical control applications in vertical domains," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 22. 104, Jun. 2021, version 18.1.0.

[2] A. Aijaz, "Private 5G: The future of industrial wireless," *IEEE Ind. Electron. Mag.*, vol. 14, no. 4, pp. 136–145, Dec. 2020.

[3] K. Zanbouri, M. Noor-A-Rahim, J. John, C.J. Sreenan, H. V. Poor, and D. Pesch, "A comprehensive survey of wireless time-sensitive networking (TSN): Architecture, technologies, applications, and open issues," *IEEE Commun. Surveys Tuts.*, vol. 27, no. 4, pp. 2129–2155, Aug. 2025.

[4] T. Zhang, C. Xue, J. Wang, Z. Yun, N. Lin, and S. Han, "A survey on industrial Internet of Things (IIoT) testbeds for connectivity research," 2024, *arXiv:2404.17485*.

[5] T. Zhang, G. Wang, C. Xue, J. Wang, M. Nixon, and S. Han, "Time-sensitive networking (TSN) for industrial automation: Current advances and future directions," *ACM Comput. Surv.*, vol. 57, no. 2, pp. 1–38, 2024.

[6] S. Petersen and S. Carlsen, "WirelessHART versus ISA100.11a: The format war hits the factory floor," *IEEE Ind. Electron. Mag.*, vol. 5, no. 4, pp. 23–34, Dec. 2011.

[7] R. Steigmann and J. Endresen, "Introduction to WISA: WISA-wireless interface for sensors and actuators," White paper, ABB, 2006.

[8] D. Dujovne, T. Watteyne, X. Vilajosana, and P. Thubert, "6TiSCH: Deterministic IP-enabled industrial internet (of things)," *IEEE Commun. Mag.*, vol. 52, no. 12, pp. 36–41, Dec. 2014.

[9] V. P. Modekurthy, A. Saifullah, and S. Madria, "DistributedHART: A distributed real-time scheduling system for WirelessHART networks," in *Proc. IEEE Real-Time Embedded Technol. Appl. Symp.*, 2019, pp. 216–227.

[10] 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 *Proc. IEEE Real-Time Embedded Technol. Appl. Symp.*, 2018, pp. 1–12.

[11] J. Wang, T. Zhang, D. Shen, X. S. Hu, and S. Han, "APaS: An adaptive partition-based scheduling framework for 6TiSCH networks," in *Proc. IEEE Real-Time Embedded Technol. Appl. Symp.*, 2021, pp. 320–332.

[12] J. Wang, Y. Liu, S. Niu, and H. Song, "Reinforcement learning optimized throughput for 5G enhanced swarm UAS networking," in *Proc. IEEE Int. Conf. Commun.*, 2021, pp. 1–6.

[13] Y. Chen, Y. Wu, Y. T. Hou, and W. Lou, "mCore: Achieving sub-millisecond scheduling for 5G MU-MIMO systems," in *Proc. IEEE Conf. Comput. Commun.*, 2021, pp. 1–10.

[14] 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 *Proc. IEEE Conf. Comput. Commun.*, 2022, pp. 130–139.

[15] E. Fountoulakis, T. Charalambous, A. Ephremides, and N. Pappas, "Scheduling policies for AoI minimization with timely throughput constraints," *IEEE Trans. Commun.*, vol. 71, no. 7, pp. 3905–3917, Jul. 2023.

[16] 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 *Proc. IEEE 39th Int. Conf. Distrib. Comput. Syst.*, 2019, pp. 1466–1475.

[17] C. Li et al., "Minimizing AoI in a 5G-based IoT network under varying channel conditions," *IEEE Internet Things J.*, vol. 8, no. 19, pp. 14543–14558, Oct. 2021.

[18] E. Ghoreishi, B. Abolhassani, Y. Huang, S. Acharya, W. Lou, and Y. T. Hou, "Cyrus: A DRL-based puncturing solution to URLLC/eMBB multiplexing in O-RAN," in *Proc. 33rd Int. Conf. Comput. Commun. Netw.*, 2024, pp. 1–9.

[19] D. Shen, T. Zhang, J. Wang, Q. Deng, S. Han, and X. S. Hu, "QoS guaranteed resource allocation for coexisting eMBB and URLLC traffic in 5G industrial networks," in *Proc. IEEE 28th Int. Conf. Embedded Real-Time Comput. Syst. Appl.*, 2022, pp. 81–90.

[20] A. Mamane, M. Fattah, M. El Ghazi, M. El Bekkali, Y. Balboul, and S. Mazer, "Scheduling algorithms for 5G networks and beyond: Classification and survey," *IEEE Access*, vol. 10, pp. 51643–51661, 2022.

[21] A. Nota, S. Saidi, D. Overbeck, F. Kurtz, and C. Wietfeld, "Context-based latency guarantees considering channel degradation in 5G network slicing," in *Proc. IEEE Real-Time Syst. Symp.*, 2022, pp. 253–265.

[22] A. Nota, S. Saidi, D. Overbeck, F. Kurtz, and C. Wietfeld, "Providing response times guarantees for mixed-criticality network slicing in 5G," in *Proc. Des., Automat. Test Europe Conf. Exhib.*, 2022, pp. 552–555.

[23] A. Shashin, A. Belogaev, A. Krasilov, and E. Khorov, "Adaptive parameters selection for uplink grant-free URLLC transmission in 5G systems," *Comput. Netw.*, vol. 222, 2023, Art. no. 109527.

[24] Y. Pan, R. Mahfouzi, S. Samii, P. Eles, and Z. Peng, "Resource optimization with 5G configured grant scheduling for real-time applications," in *Proc. Des., Automat. Test Europe Conf. Exhib.*, 2023, pp. 1–2.

[25] T. Zhang, X. S. Hu, and S. Han, "Contention-free configured grant scheduling for 5G URLLC traffic," in *Proc. 60th ACM/IEEE Des. Automat. Conf.*, 2023, pp. 1–6.

[26] T. Zhang, J. Wang, X. S. Hu, and S. Han, "Real-time flow scheduling in industrial 5G new radio," in *Proc. IEEE Real-Time Syst. Symp.*, 2023, pp. 371–384.

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

[28] V. P. Modekurthy, A. Saifullah, and S. Madria, "A distributed real-time scheduling system for industrial wireless networks," *ACM Trans. Embedded Comput. Syst.*, vol. 20, no. 5, pp. 1–28, 2021.

[29] J. Wang, T. Zhang, X. S. Hu, and S. Han, "Resource virtualization with end-to-end timing guarantees for multi-hop multi-channel real-time wireless networks," in *Proc. IEEE Real-Time Syst. Symp.*, 2023, pp. 385–396.

[30] S. S. Nakkina, S. Balijepalli, and C. R. Murthy, "Performance benchmarking of the 5G NR PHY on the OAI codebase and USRP hardware," in *Proc. 25th Int. ITG Workshop Smart Antennas*, 2021, pp. 1–6.

[31] Y. Huang, Y. T. Hou, and W. Lou, "DELUXE: A DL-based link adaptation for URLLC/eMBB multiplexing in 5G NR," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 1, pp. 143–162, Jan. 2022.

[32] N. Nikaein, M. K. Marina, S. Manickam, A. Dawson, R. Knopp, and C. Bonnet, "Openairinterface: A flexible platform for 5G research," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 44, no. 5, pp. 33–38, 2014.

[33] 3GPP, "NR; physical layer procedures for data," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 38. 214, 2018, version 15.0.0.

[34] V. Fernández-López, K. I. Pedersen, B. Soret, J. Steiner, and P. Mogensen, "Improving dense network performance through centralized scheduling and interference coordination," *IEEE Trans. Veh. Technol.*, vol. 66, no. 5, pp. 4371–4382, May 2017.

[35] 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.*, 2018, pp. 1–6.

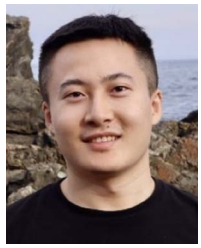
[36] 3GPP, "NR; physical channels and modulation," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 38. 211, vol. 9, 2018.

[37] M. Düngen et al., "Channel measurement campaigns for wireless industrial automation," *At-Automatisierungstechnik*, vol. 67, no. 1, pp. 7–28, 2019.

[38] 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, 2021, Art. no. 4861.

[39] J. Blazewicz, M. Y. Kovalyov, M. Machowiak, D. Trystram, and J. Weglarz, "Preemptable malleable task scheduling problem," *IEEE Trans. Comput.*, vol. 55, no. 4, pp. 486–490, Apr. 2006.
[40] H. Kanemitsu, M. Hanada, and H. Nakazato, "Clustering-based task scheduling in a large number of heterogeneous processors," *IEEE Trans. Parallel Distrib. Syst.*, vol. 27, no. 11, pp. 3144–3157, Nov. 2016.
[41] H. Chen, A. M. K. Cheng, and Y.-W. Kuo, "Assigning real-time tasks to heterogeneous processors by applying ant colony optimization," *J. Parallel Distrib. Comput.*, vol. 71, no. 1, pp. 132–142, 2011.
[42] W. Y. Lee, "Energy-saving DVFS scheduling of multiple periodic real-time tasks on multi-core processors," in *Proc. 13th IEEE/ACM Int. Symp. Distrib. Simul. Real Time Appl.*, 2009, pp. 216–223.
[43] X. Gandibleux, X. Delorme, and V. T'Kindt, "An ant colony optimisation algorithm for the set packing problem," in *Proc. Int. Workshop Ant Colony Optim. Swarm Intell.*, Springer, 2004, pp. 49–60.
[44] M. R. Garey and D. S. Johnson, *Computers and Intractability*, vol. 174. San Francisco, CA, USA: Freeman, 1979.
[45] B. Chandra and M. M. Halldórsson, "Greedy local improvement and weighted set packing approximation," *J. 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] M. Orlowski, "A new algorithm for the largest empty rectangle problem," *Algorithmica*, vol. 5, pp. 65–73, 1990.
[48] 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.*, 2020, pp. 1–6.
[49] A. A. A. Ari, A. Gueroui, C. Titouna, O. Thiare, and Z. Aliouat, "Resource allocation scheme for 5G C-RAN: A swarm intelligence based approach," *Comput. Netw.*, vol. 165, 2019, Art. no. 106957.
[50] D. Van Huynh et al., "URLLC edge networks with joint optimal user association, task offloading and resource allocation: A digital twin approach," *IEEE Trans. Commun.*, vol. 70, no. 11, pp. 7669–7682, Nov. 2022.
[51] G. Dósa and J. Sgall, "First fit bin packing: A tight analysis," in *Proc. 30th Int. Symp. Theor. Aspects Comput. Sci.*, Schloss-Dagstuhl-Leibniz Zentrum für Informatik, 2013, pp. 538–549.
[52] 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 Trans. Mobile Comput.*, vol. 20, no. 6, pp. 2111–2123, Jun. 2021.
[53] "Openairinterface 5G RAN project," (n.d.). [Online]. Available: https://gitlab.eurecom.fr/oai/openairinterface5g/-/tree/develop
[54] R.-M. Ursu, A. Papa, and W. Kellerer, "Experimental evaluation of downlink scheduling algorithms using openairinterface," in *Proc. IEEE Wireless Commun. Netw. Conf.*, 2022, pp. 84–89.
[55] 3GPP, "Technical specification group services and system aspects," 3rd Generation Partnership Project (3GPP), Technical Report (TR) 21.915, 2019, version 15.0.0.
[56] D. Liu, Y. Zhao, G. Liu, C. Wang, L. Zhou, and Y. Qian, "Real-time performance evaluation for 5G multi-link communication in industrial application," *IEEE Access*, vol. 13, pp. 26864–26875, 2025.
[57] G. C. Buttazzo and G. Buttanzo, *Hard Real-Time Computing Systems*, vol. 356. Berlin, Germany: Springer, 1997.
[58] P. Guan et al., "Ultra-low latency for 5G-A lab trial," 2016, *arXiv:1610.04362*.
[59] G. Durisi, T. Koch, and P. Popovski, "Toward massive, ultrareliable, and low-latency wireless communication with short packets," in *Proc. IEEE*, vol. 104, no. 9, pp. 1711–1726, Sep. 2016.
[60] C.-P. Li, J. Jiang, W. Chen, T. Ji, and J. Smee, "5G ultra-reliable and low-latency systems design," in *Proc. Eur. Conf. Netw. Commun.*, 2017, pp. 1–5.
[61] T. Wirth, M. Mehlhose, J. Pilz, B. Holfeld, and D. Wieruch, "5G new radio and ultra low latency applications: A PHY implementation perspective," in *Proc. 50th Asilomar Conf. Signals, Syst. Comput.*, 2016, pp. 1409–1413.
[62] F. Schaich, T. Wild, and Y. Chen, "Waveform contenders for 5G-suitability for short packet and low latency transmissions," in *Proc. IEEE 79th Veh. Technol. Conf.*, 2014, pp. 1–5.
[63] M. E. Haque, F. Tariq, M. R. Khandaker, K.-K. Wong, and Y. Zhang, "A survey of scheduling in 5G URLLC and outlook for emerging 6G systems," *IEEE access*, vol. 11, pp. 34372–34396, 2023.
[64] M. Maray, A. Jhumka, A. Chester, and M. Younis, "Scheduling dependent tasks in edge networks," in *Proc. IEEE 38th Int. Perform. Comput. Commun. Conf.*, 2019, pp. 1–4.
[65] 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.*, 2009, pp. 815–820.

**Tianyu Zhang** (Member, IEEE) received the MS and PhD degrees from Northeastern University, China, in 2013 and 2018, respectively. He was a postdoctoral associate with the University of Connecticut, working on real-time scheduling in Industrial IoT systems. He is currently an assistant professor with the Department of Computer Science at the University of Iowa, Iowa City, IA, USA. His research interests include real-time systems, cyber-physical systems, and wireless sensor networks.



**Jiachen Wang** received the BS degree from Xidian University, in 2010 and the PhD degree from the University of Connecticut, in 2024. He is currently a research scientist with Intelligent Fusion Technology, Germantown, MD, USA. His research interests include the design of real-time systems, and wireless sensor networks.

**X. Sharon Hu** (Fellow, IEEE) received the BS degree from Tianjin University, Tianjin, China, in 1982, the MS degree from the Polytechnic Institute of New York, Brooklyn, NY, USA, in 1984, and the PhD degree from Purdue University, West Lafayette, IN, USA, in 1989. She is currently the Leo E. and Patti Ruth Linbeck professor of engineering with the Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN, USA. Her research interests include energy/reliability-aware system design, circuit and architecture design with emerging technologies, real-time embedded systems, and hardware-software co-design. She has published more than 500 papers in these areas.

**Song Han** (Member, IEEE) received the BS degree in computer science from Nanjing University, Nanjing, China, in 2003, the MPhil degree in computer science from the City University of Hong Kong, Hong Kong, in 2006, and the PhD degree in computer science from the University of Texas, Austin, TX, USA, in 2012. He is currently an associate professor and a Castleman term professor of engineering innovation with the Department of Computer Science and Engineering, University of Connecticut, Storrs, CT, USA. His research interests include cyber–physical systems, real-time and embedded systems, and wireless networks. He is an associate editor of the *ACM Transactions on Cyber-Physical Systems*.