IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 19, 2018, accepted November 5, 2018, date of publication November 9, 2018,
date of current version December 19, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2880289

5G Centralized Multi-Cell Scheduling for URLLC: Algorithms and System-Level Performance

ALI KARIMI[1], (Member, IEEE), KLAUS I. PEDERSEN[1,2], (Senior Member, IEEE), NURUL HUDA MAHMOOD[1], (Member, IEEE), JENS STEINER[2], AND PREBEN MOGENSEN[1,2], (Member, IEEE)

[1]Wireless Communications Networks Section, Department of Electronic Systems, Aalborg University, 9220 Aalborg, Denmark
[2]Nokia Bell Labs, 9220 Aalborg, Denmark

Corresponding author: Ali Karimi (alk@es.aau.dk)

This work was supported by the framework of the Horizon 2020 project ONE5G receiving funds from the European Union under Grant ICT-760809.

ABSTRACT We study centralized radio access network (C-RAN) with multi-cell scheduling algorithms to overcome the challenges for supporting ultra-reliable low-latency communications (URLLC) in the fifth-generation new radio (5G NR) networks. Low-complexity multi-cell scheduling algorithms are proposed for enhancing the URLLC performance. In comparison with the conventional distributed scheduling, we show that the C-RAN architecture can significantly reduce undesirable queuing delay of URLLC traffic. The gain of user scheduling with different metrics and the benefit of packet segmentation are analyzed. The performance of the proposed solutions is evaluated with an advanced 5G NR compliant system-level simulator with high degree of realism. The results show that the centralized multi-cell scheduling achieves up to 60% latency improvement over the traditional distributed scheduling while fulfilling the challenging reliability of URLLC. It is shown that segmentation brings additional performance gain for both centralized and distributed scheduling. The results also highlight the significant impact of channel- and delay-aware scheduling of URLLC payloads.

INDEX TERMS 5G, URLLC, packet scheduling, segmentation, scheduling metric.

I. INTRODUCTION

A. SETTING THE SCENE

The third generation partnership program (3GPP) has recently released the first specifications for the fifth generation (5G) radio system, also known as the 5G New Radio (NR) [1]. The 5G NR is designed to fulfill the IMT2020 requirements [2]–[4], being able to support a diverse set of services with different characteristics and quality-of-service (QoS) targets. One of the challenging service categories is ultra-reliable low-latency communication (URLLC), where the most stringent requirement is 1 msec one-way latency in the radio access network with 99.999% reliability. However, the 5G NR is also designed to support other classes of URLLC requirements as defined in the 5G QoS class indices (5QI) with latency budgets of, for instance 5, 10, and 20 msec, as well as reliability targets from 99% to 99.999% [5].

Meeting the URLLC requirements is obviously a challenging task, especially when considering a highly dynamic multi-cell and multi-user system. Our hypothesis is that a centralized radio access network (C-RAN) architecture with fast multi-cell scheduling is an attractive solution for improving the downlink latency of URLLC, while still fulfilling the reliability requirements. We validate this hypothesis in this paper, starting with a compact overview of previous URLLC studies, followed by further crystallization of our contributions.

B. RELATED STUDIES

A large number of URLLC related studies have been published during recent years, so it would be too exhaustive to quote all here. Hence, only some relevant examples of which are summarized in the following. Popovski *et al.* [6] discuss the principles and enablers of URLLC by considering different design aspects. A recent overview paper has been published in [7], focusing on the medium access (MAC) and physical (PHY) layer enablers considered for NR standardization to make URLLC come true. There have been numerous studies on dynamic link adaptation for URLLC in [8] and [9], diversity and coding techniques [10],

hybrid automatic repeat request (HARQ) enhancements in [11] and [12], and variable transmission time intervals (TTIs) [13], [14]. An overview of the scheduler options in 5G NR is provided in [15], including descriptions of new scheduling formats and degrees of freedom added to facilitate URLLC and other services. In [16], Liu and Bennis study the effect of power allocation for URLLC vehicle-to-vehicle transmission. Several studies also find that queuing delay is a major threat for fulfilling URLLC requirements [17], [18]. As an example, even for homogeneous macro cellular deployments with spatial uniform traffic and Poisson arrival data bursts, some cells may likely experience temporary high loads, and consequently cause queuing delays that can exceed the maximum tolerable latency.

Centralized multi-cell scheduling has been studied earlier for LTE systems with mobile broadband (MBB) traffic for improving the average user experienced data rates [19]. However, to the best of our knowledge, there are very few 5G NR studies of centralized multi-cell scheduling for URLLC use cases. The study in [20] is one such example. Numerous studies have also investigated different cell association and packet scheduling methods in wireless networks. Most of the contributions are proposed for MBB traffic, based on theoretical results and mostly with high computational complexity [21], [22]. The performance evaluation of proposed contributions on practical systems without simplified assumptions and by considering the network limitations and imperfections is still an open research area [23], [24].

C. OUR CONTRIBUTION

In the 5G era, C-RAN architectures are expected to gain further popularity, especially in areas where fiber availability is present to realize front-haul connections with practically zero latency becomes a viable option [25]. Thereby, allowing centralization of resource management procedures to overcome some of the challenges for supporting URLLC. Centralized multi-cell scheduling offers numerous benefits such as increased diversity (e.g. if using dynamic point selection [26]) and the ability to reduce queuing delays as individual users data can be flexibly scheduled from different cells, as compared to more traditional distributed network architectures where users are scheduled from their single serving cell all the time.

We build on the quoted studies and propose improved centralized multi-cell scheduling algorithms for the 5G NR to enhance the URLLC performance. The starting point for the study is a realistic system model in line with the 3GPP NR specifications, adopting the advanced performance assessment models used in 3GPP. The system model comprises a multi-cell deployment with dynamic user traffic models, three-dimensional (3D) channel propagation, the 5G NR protocol stack, flexible frame structure, scheduling, link adaptation, HARQ, MIMO transmission and reception, etc. The dynamic varying overhead from sending scheduling grants to the users is taken explicitly into account. As compared to the our earlier study in [20], enhanced multi-cell scheduling algorithms are proposed and a more detailed system-level performance assessment is presented. In our search for such algorithms, we prioritize solutions of the modest complexity that are feasible for C-RAN architecture implementations, offering additional insight on the trade-offs between achievable performance and the use of sub-optimal algorithms with acceptable complexity.

Attractive multi-cell scheduling algorithms are presented, including cases with/without segmentation of the URLLC payloads over multiple transmission opportunities. That is, without segmentation, only the full URLLC payloads of modest size 50 bytes are scheduled, while for cases with segmentation, we allow that a URLLC payload is segmented so it is transmitted over multiple TTIs. Cases without segmentation have the advantage of aiming for single-shoot transmission of URLLC payloads, at the cost of not always being able to utilize all transmission resources as there may be insufficient resources to transmit full URLLC payloads. On the contrary, use of segmentation allows better utilization of radio resources, but at the expense of (i) higher control channel overhead as each transmission is accompanied with scheduling grant, as well as (ii) possibility of errors at each transmission. The trade-offs between allowing segmentation vs no segmentation therefore signify an interesting problem, which to the best of our knowledge has not yet been fully addressed. In summary, our main contributions in this article are:

- Adopting a highly detailed 5G NR compliant system-model with detailed representation of a macro cellular environment and the many performance determining C-RAN mechanisms for studying URLLC.
- Attractive sub-optimal centralized multi-cell scheduling algorithms for enhancing the URLLC system-level performance of acceptable computational complexity, including cases with/without segmentation of URLLC payloads.
- State-of-the-art system-level performance analysis of centralized multi-cell scheduling performance for URLLC cases by means of advanced system-level simulations.

Given the complexity of the considered system-model and related scheduling problems, mainly heuristic methods are applied in deriving the proposed algorithms. The corresponding performance analysis is conducted in a dynamic multi-user, multi-cell setting with high degree of realism. Due to the complexity of the system model, we rely on advanced system-level simulations for results generation. Those simulations are based on commonly accepted mathematical models, calibrated against the 3GPP 5G NR assumptions [2], making sure that statistical reliable results are generated.

The rest of the paper is organized as follows: In Section II, we outline the system model and a more detailed problem formulation of the multi-cell scheduling challenge for URLLC. In Section III the proposed multi-cell scheduling algorithms are presented. The system-level simulation methodology

appears in Section IV, followed by performance results in Section V. Finally, the study is concluded in Section VI.

II. SYSTEM MODEL AND PROBLEM FORMULATION

In line with [17] and [20], and the 3GPP NR specifications [27], we outline the assumed system model in the following, as well as present the problem formulation in greater details.

A. NETWORK TOPOLOGY AND TRAFFIC MODEL

We consider C-RAN architecture as depicted in Fig. 1 comprises of one centralized unit (CU) controlling several remote radio heads (RRHs) in a large geographical area. Ideal lossless and zero-latency communication via fiber optic cables is assumed between the CU and RRHs. The interface between the CU and the RRHs corresponds to split option-7 [28], also known as the F2 interface that can be realized with the common public radio interface (CPRI), or the enhanced CPRI (eCPRI). In line with the 3GPP defined NR architecture (see [1] and [29]), the CU hosts all the radio access network protocols from the higher PHY and upwards. Hence, including the service data adaptation protocol (SDAP), packet data convergence protocol (PDCP), radio link control (RLC), and MAC that holds the scheduling responsibility, as well as the control plane protocol and radio resource control (RRC) functionality. Thus, the RRH only includes the lower PHY functions.

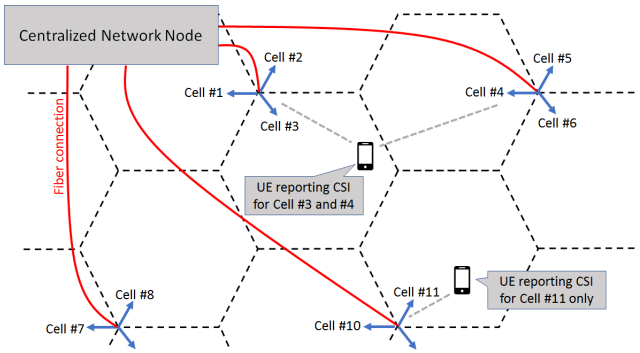

FIGURE 1. Network deployment with network elements.

The 3GPP urban macro (UMa) deployment is assumed where the RRHs are deployed in a sectorized macro cellular deployment with 500 meters inter-site distance, each hosts three sectors (cells) [2], [17]. A set of $\overline{U}$ URLLC users (UE) are randomly placed in the network area with uniform distribution. A birth-dead traffic model is assumed for each URLLC UE in which a burst of small payloads of B bytes arrive at the CU according to the Poisson distribution with an average arrival rate of λ packet per second. This traffic model is known as FTP3 in 3GPP [27]. The average offered load per cell equals to $L = 8 \cdot \overline{U} \cdot B \cdot \lambda / C$ bps/cell, where C denotes the number of cells in the network area.

B. BASIC RADIO ASSUMPTIONS

In line with [19] and [20], each UE measures the average reference symbol received power (RSRP) from the cells that it can hear and creates its channel state information (CSI) measurement set of maximum Q ($Q \geq 1$) cells it can connect to. The measurement set contains the cell with the highest received power denoted as the primary cell. It also includes up to the $Q-1$ other strongest secondary cells within the power range of W dB as compared to the primary one.

The UE measures the channel and interference for each of the cells in the CSI measurement set and reports the CSI to the network. The value of Q limits the computational complexity of CSI measurement as well as the CSI feedback overhead. Parameter W helps to control that the measurement set contain cells with sufficiently good channel quality.

Users are dynamically time-frequency multiplexed on a shared channel, using orthogonal frequency division multiple access (OFDMA). A 15 kHz sub-carrier spacing is assumed, where one physical resource block (PRB) equals 12 sub-carriers. A short TTI size of 0.143 msec, corresponding to a mini-slot of 2 OFDM symbols is assumed. The minimum scheduling resolution is one TTI (time-domain) and one PRB (frequency domain). Considering 10 and 20 MHz bandwidth (BW) configurations, the total number of available PRBs equals to $D_{total} = 50$ and $D_{total} = 100$ PRBs, respectively.

The network is only allowed to schedule a user from a cell that belongs to the user's CSI measurement set, and only from one cell per TTI. Whenever the MAC schedules a user on a certain set of resources, both a user-specific scheduling grant on the physical downlink control channel (PDCCH) and the actual transport block (data) on the physical downlink shared channel (PDSCH) are transmitted. In line with [15] and [17], the scheduling grant on the PDCCH is transmitted with aggregation levels of one to eight (or even 16) to ensure good reception quality at the UE. The data transmission on the PDSCH relies on fast link adaptation where the effective coding rate and modulation scheme is set per transmission (and communicated to the UE as part of the scheduling grant).

The link adaptation for PDCCH (i.e. setting of the aggregation level) and PDSCH is based on the received CSI from the user. As the CSI is subject to reporting delays (and other imperfections), we rely on the well-known outer loop link adaptation (OLLA) to control the block error rate (BLER). As in [8] and [17], the OLLA is set to 1% BLER for the first PDSCH transmission. If the UE fails to correctly decode a downlink scheduled data transmission, it will feed back a negative acknowledgement (NACK), and the network will later schedule a corresponding HARQ retransmission. Asynchronous HARQ is assumed for the 5G NR [11]. Conventional Chase combining [30] is assumed to combine the signals received over multiple transmissions.

C. LATENCY PROCEDURE

The downlink one-way user latency (Υ^{tot}) is defined from the time a packet arrives at the CU, until it is successfully received at the UE. If the UE decodes the packet correctly in the first transmission, the latency equals the first transmission

delay (Υ^0) expressed as:

$$\Upsilon^0 = d^0_{q,fa} + d_{cup} + d_{tx} + d_{uep}, \tag{1}$$

where $d^0_{q,fa}$ denotes the queuing and frame alignment delay of initial transmission, d_{tx} is the payload transmission time. Processing time at the CU and UE are denoted by d_{cup} and d_{uep}, respectively. If the message is erroneously decoded, the packet is subject to HARQ retransmission(s) until either it is decoded successfully or the maximum retransmissions (ϱ) is reached. In this case, Υ^{tot} can be formulated as:

$$\begin{aligned} \Upsilon^{tot} &= \Upsilon^0 + \sum_{i=1}^{r} \Upsilon^i, \\ \Upsilon^i &\triangleq d^i_{q,fa} + d^{RTT}_{HARQ}, \end{aligned} \tag{2}$$

where $r \in [1, \cdots, \varrho]$ and Υ^i denote the number of retransmissions and the i-th retransmission delay ($i \geq 1$). The HARQ round trip time is denoted by d^{RTT}_{HARQ}. In line with [17], we assume that the minimum retransmission delay is equal to $d^{RTT}_{HARQ} = 4$ TTIs.

The processing times (d_{cup} and d_{uep}) are considered to be constant with the length of 3 OFDM symbols at both the network and the receiver end [31]. The transmission time is a discrete random variable. Depending on the packet size, channel quality, and the number of assigned PRBs, d_{tx} varies from one to multiple TTIs. The frame alignment delay is a random variable with uniform distribution between 0 and 1 TTI. The queuing delay is defined as the waiting time for getting scheduled at physical layer. It is a random variable and depends on various network parameters such as the payload size, channel quality and required QoS, number of available resources, network load, and the scheduling algorithm.

It has earlier been attempted to study the effect of queuing delays by adopting multi-class queuing network models as considered [32], [33]. For such models users connected to the same cell are categorized in Γ different classes $\mathbf{k} = \{k_1, k_2, ..., k_\Gamma\}$ where members of each class share the same signal to interference plus noise ratio (SINR). On a TTI basis, the packet arrival of k-th class is modeled as a Poisson distribution with the average of $\lambda_k = \overline{u}_k \times \lambda_{TTI}$. $\overline{u}_k$ and λ_{TTI} are the number of UEs in k-th class and the user average packet arrival rate in each TTI, respectively. Note that $\overline{u}_k$ changes with channel variation. Although such models do offer some valuable insight, they fail to fully capture all performance-determining factors of the system model, and in particularly interference coupling between cells, causing random SINR fluctuations.

In a time instance, assume there are u_k UEs with pending data in k-th class, each requires r_k PRBs to transmit the packet. One or some of the UEs are subjected to queuing/multiple TTI transmission delay if

$$\sum_{k=1}^{\Gamma} u_k r_k > D_{total}.$$

D. PROBLEM FORMULATION

The CU has the following information available at each TTI:

1) Which users have pending HARQ retransmissions.
2) Which users have new data and the corresponding buffering delay.
3) From which cells the users are schedulable (i.e. corresponding to the UEs CSI measurement set).
4) An estimate of the number of PRBs for transmission of both the data and PDCCH for the cells in the CSI measurement set.

The overall objective is to maximize the tolerable average served traffic load L, while still ensuring that all payloads are delivered within a given latency budget, T_{target}, with a reliability of P_{target}, expressed as $P(\Upsilon^{tot} \leq T_{target}) \geq P_{target}$. In order to minimize the undesirable control channel overhead that unavoidable comes from segmentation of a payload over multiple TTIs, we first aim for single TTI transmission of the full URLLC payloads. For a multi-cell multi-user network of U UEs with pending data and C cells, we formulate a joint scheduling problem by defining the scheduling matrix $\mathbf{M} \in \mathbb{R}_+^{U \times C}$. Element m_{uc} of $\mathbf{M}$ is the scheduling metric for user u on cell c used for multi-cell scheduling decisions. It is assumed that $m_{uc} = 0$ for cells that are not included in the CSI measurement set of UE u. Given $\mathbf{M}$, our objective is expressed as:

$$\begin{aligned} \max_{x_{uc}} \quad & \sum_{u=1}^{U} \sum_{c=1}^{C} x_{uc} m_{uc}, \\ \text{Subject to:} \quad & \sum_{u=1}^{U} x_{uc} R_{uc} \leqslant D_{total}, \quad \forall c. \\ & \sum_{c=1}^{C} x_{uc} \leqslant 1, \quad \forall u. \\ & x_{uc} \in \{0, 1\} \quad \forall u, c, \end{aligned} \tag{3}$$

where R_{uc} denotes the estimated number PRBs to schedule UE u from cell c. Binary variable x_{uc} equals one if the u-th UE is scheduled from cell c, and otherwise zero. The first constraint is to guarantee that the summation over the number of required PRBs by the UEs associated to the same cell does not exceed total number available PRBs (D_{total}). The second constraint ensures that each UE is scheduled from at most one cell per TTI.

Note that (3) is a mixed linear integer problem which can be solved using brute-force algorithm with complexity $\mathcal{O}((Q+1)^U)$ [34]. As an example, for $U = 30$ active user in a TTI and $Q = 2$ CSIs, the complexity of optimal solution equals $3^{30} \sim 2 \times 10^{14}$. However, this is too high for practical C-RAN implementations as the scheduling decision needs to be taken every TTI and in a fast basis.

III. PROPOSED MULTI-CELL SCHEDULING

A low-complexity hierarchical joint multi-cell scheduling is proposed according to the following steps. First, pending

HARQ packets and full URLLC payloads are scheduled. Finally, segmentation is applied.

A. PENDING HARQ AND FULL PAYLOAD PACKET SCHEDULING

1) PENDING HARQ RETRANSMISSIONS

We assign the highest priority to pending HARQ retransmissions. HARQ retransmissions are scheduled immediately and from the cell which provides the best CSI. Giving the highest priority to HARQ avoids additional queuing delay of HARQ retransmissions as they are already subject to additional retransmission delay(s) of d_{HARQ}^{RTT}. Also, the probability of successful decoding increases by scheduling the UE from the cell with highest channel quality. Thus, we reduce the probability of further retransmission delays.

2) BUFFERED URLLC PACKETS

After scheduling of HARQ retransmissions, buffered packets are scheduled on the remaining PRBs. A modified matrix elimination method inspired by [19] for URLLC is adapted as follows. Based on the reported CSIs, the elements of the scheduling matrix $\mathbf{M}$ and the corresponding required number of PRBs are calculated (recall that $m_{uc} = 0$ if the c-th cell is not included in the CSI measurement set of u-th UE). If there are not enough PRBs at cell c to transmit the full payload of UE u, the corresponding scheduling metric is set to 0 meaning that UE u can not be scheduled from cell c.

At each step, the highest scheduling metric m_{uc} is selected. If there are enough PRBs at the candidate cell c to transmit the payload of UE u, the UE u is scheduled with cell c and the CU updates the number of its available PRBs as $D_c = D_c - R_{uc}$, otherwise sets $m_{uc} = 0$. To avoid user u from being co-scheduled by the other cell, the u-th row of $\mathbf{M}$ is removed. The procedure is repeated until the matrix $\mathbf{M}$ has all zero entries. The complexity of this method is $\mathcal{O}(U^3)$ [19].

A computationally efficient implementation of this method can be achieved by a sequential method as described in Algorithm 1. The approximated computational complexity of Algorithm 1 is $\mathcal{O}(Q\cdot U \log(Q\cdot U))$, while presenting the same performance as that of the matrix elimination method. It can be seen that the complexity of Algorithm 1 is significantly lower than that of the brute-force solution, making it attractive for practical C-RAN implementation.

Three different scheduling metrics are considered. Maximum throughput (Max-TP), proportional fair (PF), and throughput-delay (TP-Delay). The Max-TP aims at maximizing the achievable cell TP by prioritizing UEs reporting higher TP. In this case, the scheduling metric is defined as $m_{uc} = TP_{uc}$, where TP_{uc} is the predicted TP of the u-th UE if served by c-th cell. In line with [8], [17], and [35], we also consider the well-known proportional fair (PF) metric:

$$m_{uc} = \frac{TP_{uc}}{\overline{TP}_u}$$

where $\overline{TP}_u$ is the average delivered throughput in the past.

Algorithm 1 Proposed Algorithm for Cell Association

1: Create a vector of available PRBs at cells.
2: Schedule the HARQ transmission through the cell with the highest reported CSI and update the available number of PRBs at the serving cells.
3: For each UE that has new data, define pairs consisting of the UE and its corresponding cell candidates which the UE is schedulable.
4: Create list **s** of candidate pairs.
5: Sort candidate pairs of **s** according to the defined scheduling metric.
6: **while** Unscheduled UEs at **s** and enough PRBs at cells **do**
7: Select the first pair (u, c) of list **s**.
8: **if** $R_{uc} \leq D_c$ **then**
9: Assign UE u to cell c.
10: Update the number of available PRBs at cell c as $D_c = D_c - R_{uc}$.
11: Remove pairs corresponding to u from **s**.
12: **else**
13: Remove pair (u, c) from **s**.
14: **end if**
15: **end while**

Inspired from the well-known *Modified Largest Weighted Delay First (MLWDF)* algorithm [36], we finally define the TP-Delay metric as:

$$m_{uc} = \begin{cases} TP_{uc} & \text{if } \tau_u \leqslant 0.5 \text{ msec}, \\ \dfrac{\tau_u \cdot TP_{uc}}{\psi} & \text{if } \tau_u > 0.5 \text{ msec}, \end{cases}$$

where τ_u represents the u-th UE head of line queuing delay and ψ is equal to the time of 1 OFDMA symbol in msec. The metric increases with queuing delay and thus increases the probability of scheduling UEs with queued data.

After completion of Algorithm 1, users that can be scheduled with their full URLLC payload (one packet) have been assigned. However, there may still be some unused PRBs at some cells that could be utilized, although being insufficient to accommodate transmission of full URLLC payloads. The advantage of allowing segmentation is that higher PRB utilization is achieved, but at the cost of more generated interference because of the higher PRB utilization. Moreover, recall that to allow transmission from a cell to a UE, the available PRBs at the cell should be enough for transmission of both the PDCCH and the segmented URLLC payload at the PDSCH. The minimum required allocation size (R_{uc}^{min}) for the link between u-th UE and c-th cell is a function of the experienced SINR at the UE (obtained through the CSI). Table 1 depicts mapping of the SINR to the required number resource elements (REs) for the transmission of PDCCH and related reference signals. As the segmentation involves additional cost in terms of higher control overhead, at most one UE is segmented per cell and scheduled over remaining PRBs. Users in good channel conditions (i.e. lower control channel overhead)

TABLE 1. Mapping SINR to CCH overhead and minimum allocation size.

SINR [dB]	CCH overhead (REs)	Minimum Allocation Size (PRBs)
$[4.2, \infty)$	$1 \times 36 = 36$	4
$[0.2, 4.2)$	$2 \times 36 = 72$	6
$[-2.2, 0.2)$	$4 \times 36 = 144$	10
$(-\infty, -2.2)$	$8 \times 36 = 288$	20

are also prioritized for segmentation. Algorithm 2 is a method to allow segmentation over the cells with sufficient number of remaining PRBs (after having executed Algorithm 1), transmitting a segmented URLLC payload.

Algorithm 2 Proposed Algorithm for Segmentation

1: Create a vector of available PRBs at cells.
2: For each of the unscheduled UE, define pairs consisting of the UE and its corresponding cell candidates which have available RBs more than that of minimum required by the UE.
3: Create list **s** of candidate pairs.
4: Sort candidate pairs of **s** according to throughput.
5: **while** Unscheduled UEs at **s** and enough PRBs at cells **do**
6: Select the first pair (u, c) of list **s**.
7: **if** $R_{uc}^{min} \leq D_c$ **then**
8: Assign UE u to cell c.
9: Remove pairs corresponding to u-th UE from **s**.
10: Remove pairs corresponding to c-th cell from **s**.
11: **else**
12: Remove pair (u, c) from **s**.
13: **end if**
14: **end while**

IV. SIMULATION METHODOLOGY

The performance of the proposed algorithms is evaluated by extensive system-level simulations following the 5G NR methodology in [1] and [3]. The simulations methodology is based on commonly accepted mathematical models and is calibrated against 3GPP 5G NR assumptions [1], [2]. Table 2 summarizes the network configuration and default simulation parameters. The network operates at a carrier frequency of 2 GHz with 10 and 20 MHz bandwidth. The simulator resolution is one OFDM symbol and includes all 5G NR radio resource management functionalities outlined in Section. II.

The network consists of $C = 21$ macro cells in a three sector cellular deployment with 500 meters inter site distance. Closed-loop 2×2 single-user MIMO with rank one is assumed for all the transmissions. Each cell is configured with one panel set with $-45/+45$ degree polarization. At the UE-side, antenna polarization is $0/90$. 3GPP urban macro-3D channel model is considered [37].

A dynamic birth-death traffic model is assumed where for each UE finite-length payloads of $B = 50$ bytes are generated following a homogeneous Poisson distribution with the average of λ packet per second. Each UE performs the channel and interference estimation of the cells in the CSI measurement set periodically every 5 msec. The CSI reports are subject to 2 msec delay before being applied at the CU. In distributed scenario, each UE reports one CSI corresponding to the cell with highest RSRP value. For the centralized case, the default values of measurement set size and the window size are $Q = 2$ and $W = 10$ dB, respectively.

To suppress the noise and received interference, the UE exploits linear minimum-mean square error interference rejection combining (MMSE-IRC) receiver. After each transmission the effective SINR for each of the assigned REs is calculated and the effective exponential SINR mapping (EESM) is computed over all the scheduled RBs [38]. The calculated EESM value along with the knowledge of transmitted MCS are used to determine the probability of packet failure from detailed look-up tables that are obtained from extensive link level simulations.

The key performance indicator (KPI) for URLLC is defined as the one way achievable latency with different reliability target (i.e. 99.99%). The network URLLC capacity is defined as the maximum supported load at which the defined reliability and latency is satisfied. The simulations runs over more than 5 million packet transmissions generating results with the confidence level of 95% for the 99.999% percentile of the latency [17].

V. SIMULATION RESULTS

A. PERFORMANCE OF ALGORITHM 1

Fig. 2 depicts the complementary cumulative distribution function (CCDF) of the URLLC latency for a network with $BW = 10$ MHz bandwidth and the offered load of $L = 3.5$ Mbps/cell. The performance of the centralized Algorithm 1 is compared against that of the distributed one under different scheduling metrics. As can be seen, the centralized multi-cell scheduling significantly outperforms the distributed one. The improved latency performance is mainly

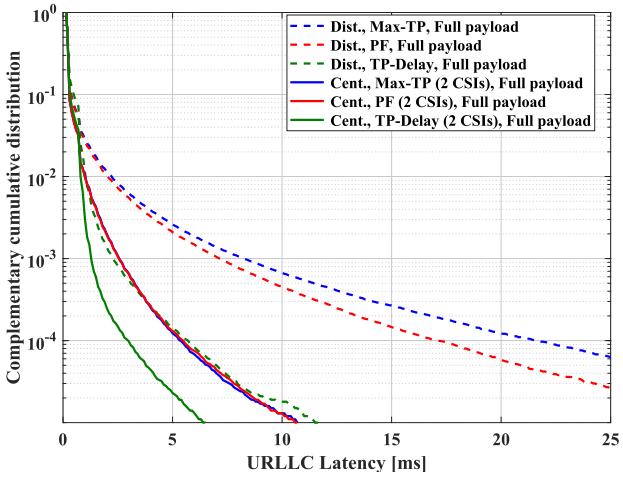

FIGURE 2. URLLC latency distribution with L = 3.5 Mbps/cell, W = 10 dB, BW=10 MHz.

TABLE 2. Default simulation assumptions.

Description	Assumption
Environment	3GPP Urban Macro (UMa); 3-sector RRHs with 500 meters inter-site distance. 21 cells.
Propagation	Urban Macro-3D
Carrier	2 GHz (FDD)
PHY numerology	15 kHz sub-carrier spacing configuration. PRB size of 12 sub-carriers (180 kHz). 24 REs in each PRB (4 REs are reserved transmission of the reference symbols. 20 REs for data) 10 and 20 MHz carrier bandwidth with 50 and 100 PRBs, respectively.
TTI sizes	0.143 msec (2-symbols mini-slot).
MIMO	Single-user 2x2 closed loop single-stream (Rank-1) configuration with $+45/-45$ cross polarization antennas at the cell, $0/90$ isotropic antenna at the UE. MMSE-IRC receiver
CSI	Periodic CSI every 5 msec, with 2 msec latency. UEs report CSI for up to Q strongest received cells that are within a power receive window of W dB. In distributed scenario, $Q = 1$ and in centralized scheduling, default is $Q = 2$ and $W = 10$ dB.
Data channel modulation and coding	QPSK to 64 QAM, with same encoding rates as specified for LTE. Turbo codes.
Link adaptation	Dynamic MCS selection with 1% initial BLER target.
HARQ	Asynchronous HARQ with Chase combining. The HARQ RTT equals minimum 4 TTIs.
User distribution	2100 (4200) URLLC UEs uniformly distributed over the network area (Average 100 (200) UEs per cell).
Traffic model	FTP3 downlink traffic with Poisson arrival of $B = 50$ bytes data bursts from each UE.
Scheduling	Max-TP, TP-Delay, PF.
Link-to-system (L2S) mapping	Based on effective exponential SINR mapping (EESM)

due to the decrease in queuing delay by exploiting available resources at secondary cells to serve more UEs. With Max-TP, the outage probability at 10^{-4} is achieved at 5.1 and 22 msec for the centralized and distributed solutions, respectively. Considering the PF metric, the latency of 17 msec for distributed solution decreases to 5.1 msec with centralized scheduling. Finally, for TP-Delay the latency improves from 5.7 msec to 3 msec. In comparison to previous studies with PF scheduling [17], [35], the TP-Delay scheduling metric provides better latency performance. At an outage probability of 10^{-4}, it achieves more than 66% and 41% latency gain under the distributed and centralized scheduling, respectively. The superior performance of the TP-Delay metric highlights the importance of channel-delay aware scheduling for URLLC. Putting the results into further perspective, it is worth noticing that end-user throughput gains of 40% from using centralized multi-cell scheduling for LTE are reported in [19] and [39] for mobile broadband file download.

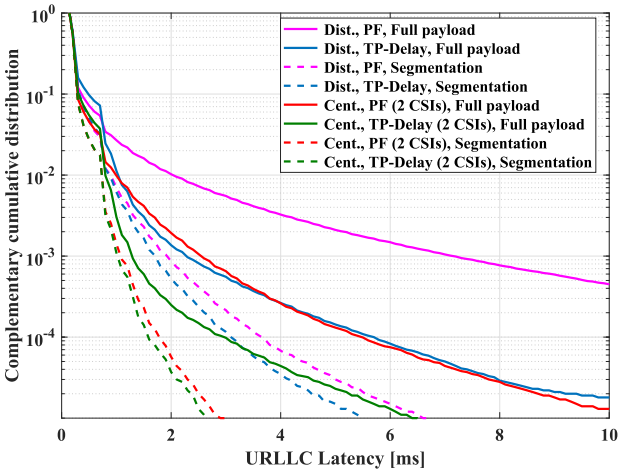

FIGURE 3. URLLC latency distribution with L = 3.5 Mbps/cell, W = 10 dB, BW = 10 MHz.

B. PERFORMANCE OF ALGORITHM 2

Now, we compare the performance of Algorithm 1 with the case where Algorithm 2 (segmentation) is also applied over the remaining PRBs after executing Algorithm 1. Figs. 3 and 4 present the CCDF of the URLLC latency for a network with 10 and 20 MHz bandwidth and different average loads of $L = 3.5$ Mbps/cell and $L = 8.5$ Mbps/cell, respectively. The results confirm that segmentation brings additional benefit for both centralized and distributed scheduling. For $BW = 10$ MHz system, it achieves significant improvements of 83% and 67% under PF for distributed and centralized scheduling. The results with TP-Delay show an improvement of 45%. The improved performance is due to the efficient utilization of all the available PRBs, thus reducing the queued data size. It is especially beneficial for low SINR UEs as they usually require large number of PRBs, which may be challenging to fit into one TTI. The main benefit of segmentation comes from applying it over the primary cells. It is usually less efficient to transmit a small part of the payload over a secondary cell as the performance degradation due to the generated interference by transmission of PDCCH becomes comparable to the achieved gain of transmitting part of the message.

Comparing 10 and 20 MHz bandwidth configuration reveals that by doubling the bandwidth, the maximum supported load that can achieve the same latency budget is more than doubled. For example, considering centralized TP-Delay scenario, 5 msec latency at the outage probability of 10^{-4} is achieved supporting $L = 4$ Mbps/cell and $L = 9.3$ Mbps/cell for 10 and 20 MHz bandwidth, respectively. Similar findings are reported in [17] and [32].

Table 3 compares the latency performances of distributed and centralized scheduling at different loads and latency

TABLE 3. Network performance for different latency budgets and at the outage probability of 10^{-4}.

Scenario	Delay [msec]					
	Bandwidth = 10 MHz			Bandwidth = 20 MHz		
Distributed, TP-Delay, Segmentation	2	5	10	2	5	10
Centralized, TP-Delay (2 CSIs), Segmentation	1.3	2.16	6	1.38	2.3	3.95
Improvement (%)	35%	57%	40%	31%	54%	60%

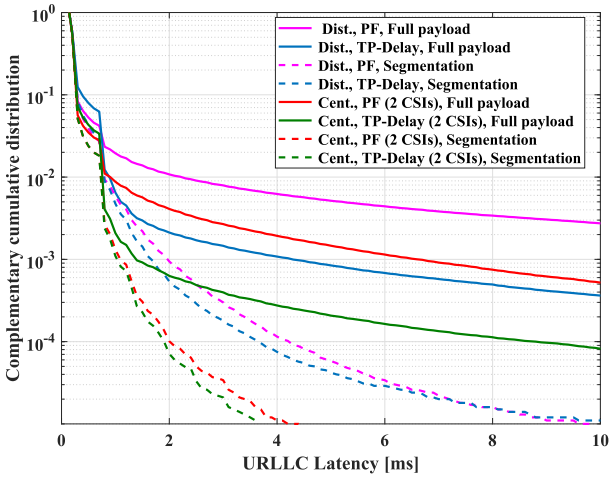

FIGURE 4. URLLC latency distribution with L = 8.5 Mbps/cell, W = 10 dB, BW = 20 MHz.

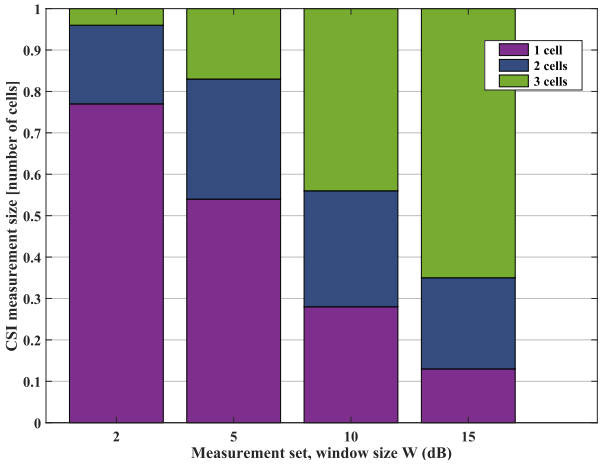

FIGURE 5. Distribution of the number of cells each user connects to, with different window size W, Q = 3, BW = 10 MHz.

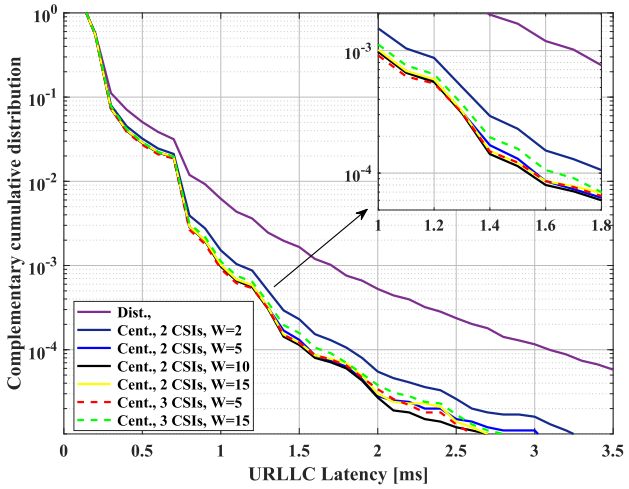

FIGURE 6. URLLC latency distribution for TP-Delay, segmentation scheduling with L = 3.5 Mbps/cell, BW = 10 MHz.

budgets at an outage probability of 10^{-4}. Centralized scheduling achieves $30\% - 60\%$ improvement with respect to that of distributed one. At low latency regimes (equivalent to low network loads), the effect of transmission delay, processing time, and HARQ RTT are dominant. As the average offered load increases, queuing delay becomes more dominant and thus the gain of centralized scheduling increases.

C. CSI MEASUREMENT SET SENSITIVITY

We next investigate the performance sensitivity versus the settings for the UEs CSI measurements (namely Q and W parameters), particularly assessing how many cells shall be considered by the centralized multi-cell scheduling algorithm for each UE. Fig. 5 illustrates the percentage of UEs having either one, two or three cells in its CSI measurement set depending on the value of W, for $Q = 3$. As expected, by increasing the value of the window size (W) the percentage of UEs with a CSI measurement size of two or three increases. For example, increasing the window size from $W = 2$ dB to $W = 15$ dB, the percentage of UEs with a CSI measurement size greater than one increases from 23% to 87%, respectively, i.e. those UEs that are subjected to multi-cell scheduling. The effect of Q and W on the URLLC performance is presented in Fig. 6.

It is interesting to note that the major improvements of the URLLC latency performance are achieved with $Q = 2$ cells and $W = 2$ dB, despite that only 23% of the UEs have a CSI measurement size of two, and thus 77% of the UEs are scheduled always from their primary cell. Increasing W to 5 dB or 10 dB results in additional performance benefits. Increasing W beyond 10 dB results in no additional gains, but rather a risk of experiencing some performance losses as cells with too weak signal strength are included in the UEs CSI measurement set. Increasing Q from 2 to 3, at the best results in minor additional benefits. The former observation partly relates to our assumption of having UEs with two receive antennas and MMSE-IRC receiver type, and thus being able to maximum suppress the interference from one dominant interfering cell. Hence, for $Q = 2$, the UE may be able to suppress the interference from its primary cell if being scheduled from its secondary cell. While if $Q = 3$, it cannot suppress the interference from both its primary cell and the strongest secondary cell, if being scheduled from the weakest secondary cell.

Fig. 7 shows the empirical CDFs of the predicted TP for the cells in the CSI measurement set for $Q = 3$, $W = 10$ dB, and different offered loads. As expected, the highest TP is observed the 1st cell (primary) where the UE receives the strongest RSRP. The supported throughput for the second and third strongest cells is clearly much lower, and hence further illustrates why the benefits of setting $Q = 3$, as compared to $Q = 2$, are marginal, and in most cases not worth considering. Hence, based on the reported findings in Figs. 5-7, we recommend using $W \in [5\ 10]$ dB and $Q = 2$. Referring to the complexity expressions for the centralized multi-cell scheduling algorithms in Section III, using $Q = 2$ (instead of $Q = 3$) also helps significantly reduce the complexity of centralized multi-cell scheduling algorithms. Similarly, the UE complexity, and uplink CSI reporting overhead is obviously more attractive for $Q = 2$, as compared to $Q = 3$.

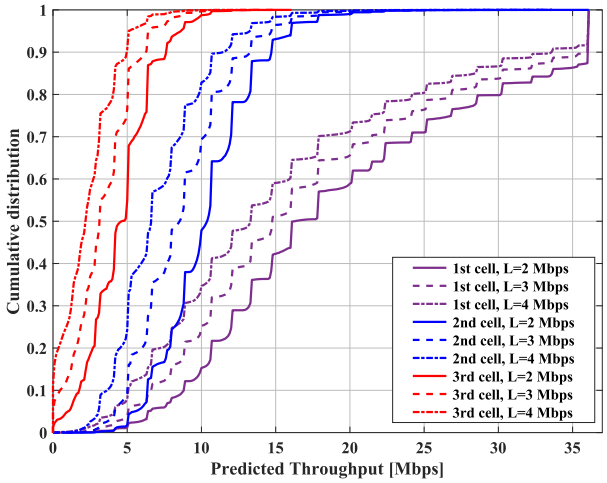

FIGURE 7. User predicted throughput of the cells in the measurement set for different load, W=10 dB, BW=10 MHz.

VI. CONCLUSION

In this paper, we have investigated centralized multi-cell scheduling of URLLC for 5G NR. Dynamic algorithms including the case with/without segmentation of URLLC payloads are proposed to improve the latency and reliability of URLLC. The solutions have low computational complexity and are attractive for practical C-RAN implementations.

The performance of the proposed solutions is evaluated by performing a variety of simulations using a highly detailed advanced 5G NR compliant system-level simulator. Results show that the proposed centralized multi-cell scheduling solutions provide significant latency performance gains of up to 60% over traditional distributed solutions. We showed that the major improvement of URLLC latency is achieved for the case with the UE CSI measurement size of $Q = 2$ cells within a power window of $W \in [5\ 10]$ dB. The results also illustrates that segmentation can reduce the queued data and bring significant URLLC latency improvement for both centralized and distributed scheduling. Finally, the importance of channel-delay aware scheduling for URLLC is shown. Future studies could examine the performance of the optimal solution, investigate more advanced interference coordination and multi-cell scheduling techniques for URLLC.

ACKNOWLEDGMENT

The authors would like to acknowledge the contributions of their colleagues in the project, although the views expressed in this contribution are those of the authors and do not necessarily represent the project.

REFERENCES

[1] *NR and NG-RAN Overall Description; Stage-2*, document 3GPP 38.300, Version 2.0.0, Dec. 2017.
[2] *Study on Scenarios and Requirements for Next Generation Access Technologies*, document 3GPP 38.913, Version 14.1.0, Mar. 2016.
[3] *IMT Vision—'Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond'*, document M.2083, ITU Radiocommunication Study Groups, Feb. 2015.
[4] E. Dahlman *et al.*, "5G wireless access: Requirements and realization," *IEEE Commun. Mag.*, vol. 52, no. 12, pp. 42–47, Dec. 2014.
[5] *System Architecture for the 5G System*, document 3GPP Technical Specification 23.501, Release 15, Dec. 2017.
[6] P. Popovski *et al.*, "Wireless access for ultra-reliable low-latency communication: Principles and building blocks," *IEEE Netw.*, vol. 32, no. 2, pp. 16–23, Mar. 2018.
[7] G. Pocovi, H. Shariatmadari, G. Berardinelli, K. Pedersen, J. Steiner, and Z. Li, "Achieving ultra-reliable low-latency communications: Challenges and envisioned system enhancements," *IEEE Netw.*, vol. 32, no. 2, pp. 8–15, Mar. 2018.
[8] G. Pocovi, B. Soret, K. I. Pedersen, and P. Mogensen, "MAC layer enhancements for ultra-reliable low-latency communications in cellular networks," in *Proc. IEEE ICC Workshops*, May 2017, pp. 1005–1010.
[9] H. Shariatmadari, Z. Li, M. A. Uusitalo, S. Iraji, and R. Jäntti, "Link adaptation design for ultra-reliable communications," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2016, pp. 1–5.
[10] D. Ohmann, A. Awada, I. Viering, M. Simsek, and G. P. Fettweis, "Diversity trade-offs and joint coding schemes for highly reliable wireless transmissions," in *Proc. IEEE 84th Veh. Technol. Conf. (VTC-Fall)*, Sep. 2016, pp. 1–6.
[11] K. I. Pedersen, S. R. Khosravirad, G. Berardinelli, and F. Frederiksen, "Rethink hybrid automatic repeat request design for 5G: Five configurable enhancements," *IEEE Wireless Commun.*, vol. 24, no. 6, pp. 154–160, Dec. 2017.
[12] G. Berardinelli, S. R. Khosravirad, K. I. Pedersen, F. Frederiksen, and P. Mogensen, "On the benefits of early HARQ feedback with non-ideal prediction in 5G networks," in *Proc. Int. Symp. Wireless Commun. Syst. (ISWCS)*, Sep. 2016, pp. 11–15.
[13] Q. Liao, P. Baracca, D. Lopez-Perez, and L. G. Giordano, "Resource scheduling for mixed traffic types with scalable TTI in dynamic TDD systems," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2016, pp. 1–7.
[14] K. I. Pedersen, M. Niparko, J. Steiner, J. Oszmianski, L. Mudolo, and S. R. Khosravirad, "System level analysis of dynamic user-centric scheduling for a flexible 5G design," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2016, pp. 1–6.
[15] K. Pedersen, G. Pocovi, J. Steiner, and A. Maeder, "Agile 5G scheduler for improved E2E performance and flexibility for different network implementations," *IEEE Commun. Mag.*, vol. 56, no. 3, pp. 210–217, Mar. 2018.
[16] C.-F. Liu and M. Bennis, "Ultra-reliable and low-latency vehicular transmission: An extreme value theory approach," *IEEE Commun. Lett.*, vol. 22, no. 6, pp. 1292–1295, Jun. 2018.
[17] G. Pocovi, K. I. Pedersen, and P. Mogensen, "Joint link adaptation and scheduling for 5G ultra-reliable low-latency communications," *IEEE Access*, vol. 6, pp. 28912–28922, 2018.
[18] E. Khorov, A. Krasilov, and A. Malyshev, "Radio resource and traffic management for ultra-reliable low latency communications," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2018, pp. 1–6.
[19] 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.

[20] A. Karimi, K. I. Pedersen, N. H. Mahmood, J. Steiner, and P. Mogensen, "Centralized joint cell selection and scheduling for improved URLLC performance," presented at the 29th Annu. IEEE Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC), Sep. 2018.
[21] J. Ghimire and C. Rosenberg, "Resource allocation, transmission coordination and user association in heterogeneous networks: A flow-based unified approach," *IEEE Trans. Wireless Commun.*, vol. 12, no. 3, pp. 1340–1351, Mar. 2013.
[22] H. Zhou, S. Mao, and P. Agrawal, "Approximation algorithms for cell association and scheduling in femtocell networks," *IEEE Trans. Emerg. Topics Comput.*, vol. 3, no. 3, pp. 432–443, Sep. 2015.
[23] D. Liu *et al.*, "User association in 5G networks: A survey and an outlook," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1018–1044, 2nd Quart. 2016.
[24] Z. Qi, T. Peng, and W. Wang, "Distributed resource scheduling based on potential game in dense cellular network," *IEEE Access*, vol. 6, pp. 9875–9886, 2018.
[25] T. Pfeiffer, "Next generation mobile fronthaul and midhaul architectures [invited]," *J. Opt. Commun. Netw.*, vol. 7, no. 11, pp. B38–B45, Nov. 2015.
[26] R. Agrawal, A. Bedekar, R. Gupta, S. Kalyanasundaram, H. Kroener, and B. Natarajan, "Dynamic point selection for LTE-advanced: Algorithms and performance," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2014, pp. 1392–1397.
[27] *Study on New Radio Access Technology Physical Layer Aspects*, document 3GPP 38.802, Version 14.0.0, Mar. 2017.
[28] *Study on New Radio Access Technology; Radio Interface Protocol Aspects*, document 3GPP 38.804, Version 14.0.0, Mar. 2017.
[29] *Technical Specification Group Radio Access Network; NG-RAN; Architecture Description*, document 3GPP 38.401, Version 15.1.0, Mar. 2018.
[30] D. Chase, "Code combining—A maximum-likelihood decoding approach for combining an arbitrary number of noisy packets," *IEEE Trans. Commun.*, vol. 33, no. 5, pp. 385–393, May 1985.
[31] *Evolved Universal Terrestrial Radio Access; Physical Layer Procedures*, document 3GPP Technical Specification 36.213, Version 15.1.0, Mar. 2018.
[32] 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. (EuCNC)*, Jun. 2017, pp. 1–5.
[33] A. Anand and G. de Veciana. (Apr. 2018). "Resource allocation and HARQ optimization for URLLC traffic in 5G wireless networks." [Online]. Available: https://arxiv.org/abs/1804.09201
[34] D. Liu, Y. Chen, K. K. Chai, T. Zhang, and M. Elkashlan, "Opportunistic user association for multi-service HetNets using Nash bargaining solution," *IEEE Commun. Lett.*, vol. 18, no. 3, pp. 463–466, Mar. 2014.
[35] A. A. Esswie and K. I. Pedersen, "Opportunistic spatial preemptive scheduling for URLLC and eMBB coexistence in multi-user 5G networks," *IEEE Access*, vol. 6, pp. 38451–38463, 2018.
[36] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, and R. Vijayakumar, "Providing quality of service over a shared wireless link," *IEEE Commun. Mag.*, vol. 39, no. 2, pp. 150–154, Feb. 2001.
[37] *Further Advancements for E-UTRA Physical Layer Aspects*, document 3GPP 36.814, Version 9.2.0, Mar. 2017.
[38] S. N. Donthi and N. B. Mehta, "An accurate model for EESM and its application to analysis of CQI feedback schemes and scheduling in LTE," *IEEE Trans. Wireless Commun.*, vol. 10, no. 10, pp. 3436–3448, Oct. 2011.
[39] V. Fernandez-Lopez, B. Soret, and K. I. Pedersen, "Joint cell assignment and scheduling for centralized baseband architectures," in *Proc. IEEE 81st Veh. Technol. Conf. (VTC Spring)*, May 2015, pp. 1–5.

ALI KARIMI received the B.Sc. degree in electrical engineering from the Isfahan University of Technology and the M.Sc. degree in communication systems from Tehran University. He is currently pursuing the Ph.D. degree with the Department of Electronic Systems, Aalborg University, in collaboration with Nokia Bell Labs, Aalborg, Denmark. His research interests include network optimization, 5G new radio, and ultra-reliable low-latency communications.

KLAUS I. PEDERSEN received the M.Sc. degree in electrical engineering and the Ph.D. degree from Aalborg University, Aalborg, Denmark, in 1996 and 2000, respectively. He is currently leading the Nokia Bell Labs Research Team, Aalborg, and a part-time Professor with the Wireless Communications Network Section, Aalborg University. He has authored/co-authored approximately 160 peer-reviewed publications on a wide range of topics, as well as an inventor on several patents. He is currently part of the EU funded research project ONE5G that focus on E2E-aware optimizations and advancements for the network edge of 5G new radio. His current work is related to 5G new radio, including radio resource management aspects, and the continued long-term evolution and its future development, with a special emphasis on mechanisms that offer improved end-to-end (E2E) performance delivery.

NURUL HUDA MAHMOOD (S'06–M'13) was born in Chittagong, Bangladesh. He received the Ph.D. degree in wireless communications from the Norwegian University of Science and Technology, Norway, in 2012. He has been with the Wireless Communication Networks Section, Aalborg University, since 2012. He is also contributing to the EU funded research project ONE5G. He has authored/co-authored over 50 peer-reviewed publications. His current research interests include the resource optimization algorithm design for URLLC services and the modeling and performance analysis of wireless communication systems.

JENS STEINER received the M.Sc. degree in electrical engineering from Aalborg University, Denmark, in 1996, with speciality in software engineering. Since 1996, he has been working for different companies mainly in the telecommunications sector. Since 2005, he has been with Nokia Bell Labs, Aalborg, first as an external consultant and subsequently as a permanent Member of Staff, where he is currently involved in 5G radio access network system-level simulator research and development. He also contributes to radio research beyond software development. He is part of the EU funded research project ONE5G that focuses on the development of a new multi-service capable 5G radio for below 6-GHz operation.

PREBEN MOGENSEN received the M.Sc. and Ph.D. degrees from Aalborg University in 1988 and 1996, respectively. He is currently a Principal Scientist at Nokia Bell Labs, Aalborg, Denmark, and a Bell Labs Fellow. He is also a Professor at Aalborg University and the Head of the Wireless Communication Networks Section. He is currently involved in research and standardization for vertical use cases for LTE and 5G, including LPWA IoT, URLLC, I.4.0, V2X, UAV, and train communication. He has published more than 400 papers within wireless communication, and he has over 19 000 Google Scholar citations.