Opportunistic Spatial Preemptive Scheduling for URLLC and eMBB Coexistence in Multi-User 5G Networks

Abdul-Mawgood Ali Ali Esswie, Ali; Pedersen, Klaus I.

Published in:
IEEE Access

DOI (link to publication from Publisher):
10.1109/ACCESS.2018.2854292

Publication date:
2018

Document Version
Publisher's PDF, also known as Version of record

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 28, 2018, accepted July 4, 2018, date of publication July 9, 2018, date of current version July 30, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2854292

Opportunistic Spatial Preemptive Scheduling for URLLC and eMBB Coexistence in Multi-User 5G Networks

ALI A. ESSWIE, (Member, IEEE), AND KLAUS I. PEDERSEN, (Senior Member, IEEE)
Nokia Bell Labs, 9220 Aalborg, Denmark
Department of Electronic Systems, Aalborg University, 9100 Aalborg, Denmark

Corresponding author: Ali A. Esswie (ali.esswie@nokia-bell-labs.com)

This work was supported in part by the Innovation Fund Denmark (IFD) under Grant 7038-00009B and in part by the Horizon 2020 Project ONE5G through the European Union under Grant ICT-760809.

ABSTRACT The fifth generation (5G) of the mobile networks is envisioned to feature two major service classes: ultra-reliable low-latency communications (URLLC) and enhanced mobile broadband (eMBB). URLLC applications require a stringent one-way radio latency of 1 ms with 99.999% success probability while eMBB services demand extreme data rates. The coexistence of the URLLC and eMBB quality of service (QoS) on the same radio spectrum leads to a challenging scheduling optimization problem, that is vastly different from that of the current cellular technology. This calls for the novel scheduling solutions which cross-optimize the system performance on a user-centric, instead of network-centric basis. In this paper, a null-space-based spatial preemptive scheduler for joint URLLC and eMBB traffic is proposed for the densely populated 5G networks. Proposed scheduler framework seeks for cross-objective optimization, where the critical URLLC QoS is guaranteed while extracting the maximum possible eMBB ergodic capacity. It utilizes the system spatial degrees of freedom in order to instantly offer an interference-free subspace for the critical URLLC traffic. Thus, a sufficient URLLC decoding ability is always preserved, and with the minimal impact on the eMBB performance. Analytical analysis and extensive system level simulations are conducted to evaluate the performance of the proposed scheduler against the state-of-the-art scheduler proposals from industry and academia. Simulation results show that the proposed scheduler offers extremely robust URLLC latency performance with a significantly improved ergodic capacity.

INDEX TERMS 5G, radio resource management, scheduling, ultra-reliable low-latency communications (URLLC), enhanced mobile broadband (eMBB), MU-MIMO, preemptive, null space.

I. INTRODUCTION

The 3rd generation partnership project (3GPP) is progressing the standardization of the fifth generation (5G) standards with a big momentum [1]–[4]. The first 5G specifications support two major service classes: ultra-reliable low-latency communications (URLLC) and enhanced mobile broadband (eMBB) [5], respectively. The URLLC denote the future applications which demand extremely reliable and low latency radio transmissions, i.e., one-way radio latency of 1 ms, associated with $1 - 10^{-5}$ success probability [6], [7]. That is, a URLLC packet is of no-use if it can not be successfully decoded within the 1 ms latency deadline. Accordingly, supporting such stringent URLLC latency specifications enables many novel use cases [8], including smart grids, tactile internet, wireless industrial control, and real time vehicle-to-vehicle communications.

However, due to the limited available spectrum in the centimeter-wave region, both eMBB and URLLC applications shall coexist on the same carrier. Thus, achieving such extreme spectral efficiency (SE) for eMBB applications and the ultra reliability and low latency for URLLC services becomes a challenging scheduling task, due to the fundamental trade-off between latency, reliability and SE [9]. For instance, to satisfy such unprecedented URLLC requirements, the system should be forcibly engineered such that blocking URLLC packets is a rare event. This can be achieved by setting an extremely tight block error rate (BLER) to preserve a sufficient URLLC signal-to-interference-noise-ratio (SINR) [10]. Consequently, URLLC users must fulfill their outage capacity of interest [11] at the expense of the overall ergodic capacity, leading to a severe loss of the network SE.

A. STATE OF THE ART URLLC SCHEDULING STUDIES

Recently, the multiplexing of coexistent URLLC and eMBB traffic on the same radio spectrum is gaining progressive research attention in both industry and academia. The agile 5G frame structure design is shown to be of great significance to satisfy the URLLC latency [12]–[15], where users can be scheduled on transmission time intervals (TTIs) of different durations. For instance, eMBB traffic is scheduled with a long TTI duration to meet its extreme SE requirements while URLLC traffic can be scheduled on a shorter TTI duration for its tight latency deadline. Nevertheless, the latter case induces an increased control signaling overhead, which in turn degrades the control channel (CCH) capacity.

Moreover, spatial diversity techniques are considered as enablers for the URLLC by preserving a sufficient received SINR point. The study in [16] demonstrates that a 4×4 multi-input multi-output (MIMO) microscopic diversity along with two orders of macroscopic diversity are essential to reach the outage SINR point, required to achieve the URLLC latency limit at the 10^{-5} outage in 3GPP macro networks. These conclusions are also supported by URLLC realistic measurement campaigns [17]. Hence, the URLLC latency budget can be achieved by enhancing the decoding ability.

The recent work in [18] further broadens the adoption of the spatial diversity for URLLC communications. It flexibly assigns different coded segments of the URLLC payload to several active interfaces, i.e., transmitters, based on the associated latency, reliability, and bit rate properties. This is a substitute of transmitting duplicate versions of the URLLC packets from different transmitters at the same time. Thus, a better latency-reliability trade-off can be achieved by reducing the original payload transmission time. Additionally, the work in [19] considers a semi-shared resource allocation algorithm for the URLLC-type communications. It avoids preserving an exclusive set of the radio resources for the URLLC traffic due to its sporadic nature; however, it splits the URLLC resource allocation into two chunks as: 1) shared resources with other eMBB traffic, and 2) dedicated single-user (SU) resources. The overall SE is enhanced; yet, with employing non-linear transceivers to compensate for the inter-user interference across the shared resources.

Furthermore, system-level packet duplication (PD) with the dual connectivity architecture in the 5G new radio (NR) [20], where users are simultaneously connected to a primary and secondary cell, is envisioned to offer great reliability levels to address such URLLC outage requirements. However, in order not to excessively consume the radio resources by redundant packets, the benefit of the URLLC PD is relevant to specific scenarios, where channels are highly unfavorable.

Additionally, the study in [21] reports advanced scheduling enhancements for optimized URLLC latency performance, including dynamic and load-dependent BLER optimization, refined hybrid automatic repeat request (HARQ) and link adaptation filtering in partly loaded cells. On another side, punctured scheduling (PS) [22] is a state-of-the-art study which aims at eliminating the scheduling queuing delay component of the stochastic URLLC traffic. If URLLC queuing is foreseen, due to resource shortage, PS scheduler instantly overwrites part of the ongoing eMBB transmissions for immediate URLLC scheduling, at the expense of a highly degraded eMBB SE. Subsequently, enhanced PS (E-PS) scheduler [23] is recently introduced to provide an improved ergodic capacity by informing the victim eMBB users of which physical resource blocks (PRBs) have been punctured by URLLC transmissions, in order to avoid erroneous Chase combing HARQ process, i.e., punctured resources are considered information-less. Code-block (CB) based HARQ re-transmission [24], [25] schemes are also proposed to reduce the overhead size of the punctured eMBB re-transmissions; however, a multi-bit HARQ ACK/NACK is required.

Finally, a multi-user-punctured scheduler (MU-PS) [26] is recently demonstrated to offer an attractive tradeoff between system ergodic capacity and URLLC (outage) performance. MU-PS first attempts to fit the sporadically incoming URLLC traffic within an ongoing eMBB traffic in a standard MU-MIMO transmission. If the MU pairing can not be satisfied at an arbitrary TTI, MU-PS scheduler falls back to PS scheduler for instant URLLC scheduling without queuing. Despite the achievable enhanced SE, MU-PS has shown a non-robust URLLC latency performance since the standard MU pairing constraint is only dependent on the rate maximization. Thus, it may lead to a further degraded SINR level of the URLLC traffic, due to the power sharing and the resulting inter-user interference.

Compared to the state-of-the-art schedulers, the URLLC outage capacity is monotonically satisfied, only with the associated dedicated resource allocation size or the provided decoding SINR level. When eMBB and URLLC traffic coexists on same spectrum, such approach results in severe degradation of the overall SE. Needless to say, a flexible scheduling framework for cross-objective optimization is still critical in scenarios where an efficient multiplexing of the eMBB and URLLC traffic is mandated.

B. PAPER CONTRIBUTION

In this work, we propose a null-space-based preemptive scheduler (NSBPS) for densely populated 5G networks. The proposed NSBPS aims to dynamically cross optimize a jointly constrained system utility, where the URLLC quality of service (QoS) is always guaranteed while achieving the maximum possible ergodic capacity. If the instantaneous schedulable radio resources are not sufficient to contain the incoming URLLC traffic, NSBPS scheduler forcibly fits the URLLC traffic within an ongoing eMBB transmission in a controlled, biased, and semi-transparent MU-MIMO transmission. Proposed scheduler pre-defines a reference spatial subspace, pointing to an arbitrary direction. Then, it instantly searches for an active eMBB transmission which is most aligned within the reference subspace. Next, NSBPS scheduler spatially projects the selected eMBB transmission onto the reference subspace, in order for its paired URLLC user

to orient its decoding vector within one possible null-space, thus, no residual inter-user interference is experienced at the URLLC user. Compared to the state-of-the-art scheduling studies from industry and academia, proposed NSBPS shows extreme robustness of the URLLC QoS with significantly enhanced ergodic capacity. The major framework of this work is summarized as follows:

- We extend our recent studies [11], [26] to propose a comprehensive performance analysis of the NSBPS scheduler under diversity of traffic and network settings.
- Compared to the state-of-the-art scheduler proposals from latest 3GPP standards, the derived NSBPS scheduler shows extreme URLLC latency robustness while approaching the network ergodic capacity.
- Proposed NSBPS scheduler is compliant with the 5G-NR standardization and requires neither excessive control overhead nor higher processing complexity.

Due to the complexity of the 5G-NR and addressed problems therein [1]–[3], the performance of the proposed NSBPS scheduler is evaluated by highly-detailed system level simulations (SLSs), and supported by analytical analysis of the key performance indicators. Following the same methodology as in [11] and [26], these simulations are based on widely accepted mathematical models and calibrated against the 3GPP 5G-NR assumptions of the majority of the resource management functionalities, e.g., HARQ, link-to-system mapping, and adaptive link adaptation. Furthermore, simulation results are ensured to be statistically reliable by preserving an extremely sufficient simulation confidence interval.

Notations: $(\mathcal{X})^{\mathrm{T}}$, $(\mathcal{X})^{\mathrm{H}}$ and $(\mathcal{X})^{-1}$ stand for the transpose, Hermitian, and inverse operations of $\mathcal{X}$, $\mathcal{X}\cdot\mathcal{Y}$ is the dot product of $\mathcal{X}$ and $\mathcal{Y}$, while $\overline{\mathcal{X}}$ and $\|\mathcal{X}\|$ represent the mean and 2-norm of $\mathcal{X}$. $\mathcal{X} \sim \mathbb{CN}(0,\sigma^2)$ presents a complex Gaussian random variable with zero mean and variance σ^2, $\mathcal{X}^{\kappa}$, $\kappa\in\{\mathrm{llc},\mathrm{mbb}\}$ denotes the type of user $\mathcal{X}$, $\mathbb{E}\{\mathcal{X}\}$ and **card**$(\mathcal{X})$ are the statistical expectation and cardinality of $\mathcal{X}$.

The rest of this paper is organized as follows. Section II introduces the system and signal models. Section III presents the addressed problem formulation. Section IV discusses the proposed NSBPS scheduler in detail. Section V describes an analytical gain analysis compared to the state-of-the-art studies, and extensive system level performance evaluation is drawn in Section VI. Section VII concludes the paper.

II. SETTING THE SCENE

A. SYSTEM MODEL

We consider a downlink (DL) 5G-NR network where the URLLC and eMBB service classes coexist [11], [26]. There are C cells, each equipped with N_t transmit antennas, and K uniformly-distributed user equipment's (UEs) per cell, each equipped with M_r receive antennas. Users are dynamically multiplexed by the orthogonal frequency division multiple access (OFDMA) [27]. We assess three types of DL traffic as: (1) URLLC sporadic FTP3 traffic with finite $\mathrm{B_{llc}}-$byte payload size and Poisson arrival process λ, (2) eMBB full buffer traffic model with infinite payload size, and (3) eMBB constant bit rate (CBR) traffic model [28], i.e., broadband video streaming, with a predetermined number of packets $\check{n}$, each is $\mathrm{B_{mbb}}-$byte, and packet inter-arrival rate $\check{\imath}$.

The average number of UEs per cell is expressed as: $K_{\mathrm{mbb}}+K_{\mathrm{llc}}=K$, where K_{mbb} and K_{llc} are the average numbers of eMBB and URLLC UEs per cell, respectively. Hence, the offered URLLC load per cell is given by: $K_{\mathrm{llc}}\times \mathrm{B_{llc}}\times\lambda$, while the eMBB full buffer load is infinite and the CBR load per cell is: $K_{\mathrm{mbb}}\times\left(\frac{\mathrm{B_{mbb}}}{(\check{n}-1)\check{\imath}}\right)$, respectively. The flexible frame structure of the 5G-NR is adopted in this work [12], where the URLLC and eMBB UEs are scheduled with variable TTI duration. As depicted in Fig. 1, eMBB traffic is scheduled per a long TTI of 14-OFDM symbols for maximizing its perceived SE while the URLLC traffic is scheduled per a shorter TTI of 2-OFDM symbols, i.e., mini-slot, due to its latency requirements. In the frequency domain, the minimum schedulable unit is the PRB, each is 12 sub-carriers of 15 kHz spacing. In line with [12] and [13], the scheduling grant is transmitted within the resources assigned to each user, i.e., in-resource CCH. Thus, the minimum resource allocation per user should be sufficiently large to accommodate the in-resource CCH in addition to its desired payload.

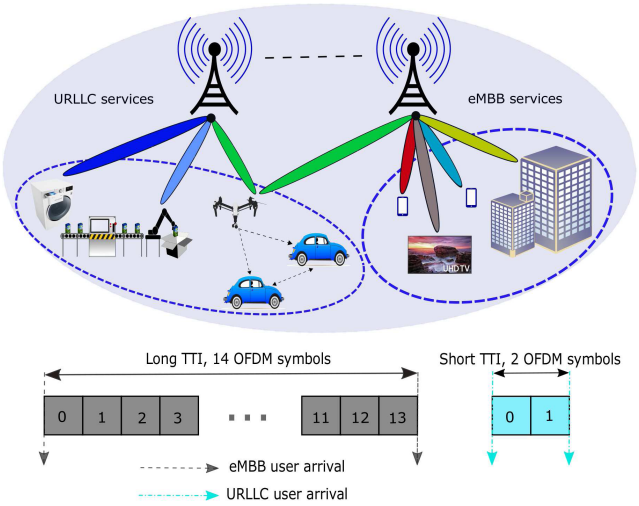

FIGURE 1. Agile 5G system model and frame structure.

Dynamic link adaptation with adaptive selection of the modulation and coding schemes (MCS) is assumed [29], based on the frequency-selective channel quality indication (CQI) user reports. Due to the bursty nature of the FTP3 URLLC and CBR eMBB traffic, the set of active interferers in the system changes sporadically in return, leading to a highly varying interference pattern. Thus, a sliding low pass filter is applied on the instantaneous CQI reports [21] to smooth out the variance of the interference pattern as

$$\partial(t)=\tilde{a}A+(1-\tilde{a})\partial(t-1), \tag{1}$$

where $\partial(t)$ is the final CQI value based on the averaged interference covariance, to be considered for MCS selection

at the t^{th} TTI, A is the CQI value calculated based on the instantaneous interference pattern, and $\tilde{a} \leq 1$ is the filter coefficient to indicate how much confidence should be given to the current reported CQI value. Finally, the Chase combining HARQ re-transmissions [30] are implemented to relax the target BLER transmission requirements, upon the reception of an associated NACK feedback.

B. SIGNAL MODEL

A MU-MIMO signal modeling is adopted in this work, where a maximum subset of MU co-scheduled URLLC-eMBB user pairs $\mathbb{G}_c \in \mathcal{K}_c$ is allowed, where $G_c = \mathbf{card}(\mathbb{G}_c)$, $G_c \leq N_t$ is the number of co-scheduled users and $\mathcal{K}_c$ is the set of active UEs in the c^{th} cell. Thus, the DL signal, received by the k^{th} user from the c^{th} cell is given by

$$\mathrm{y}^{\kappa}_{k,c} = \mathbf{H}^{\kappa}_{k,c}\mathbf{v}^{\kappa}_{k,c}s^{\kappa}_{k,c} + \sum_{g\in\mathbb{G}_c, g\neq k} \mathbf{H}^{\kappa}_{k,c}\mathbf{v}_{g,c}s_{g,c} + \sum_{j=1,j\neq c}^{C}\sum_{g\in\mathbb{G}_j}\mathbf{H}_{g,j}\mathbf{v}_{g,j}s_{g,j} + \mathbf{n}^{\kappa}_{k,c}, \quad (2)$$

where $\mathbf{H}^{\kappa}_{k,c} \in \mathcal{C}^{M_r\times N_t}, \forall k \in \{1,\ldots,K\}, \forall c \in \{1,\ldots,C\}$ is the 3D channel seen at the k^{th} user from the c^{th} cell, $\mathbf{v}^{\kappa}_{k,c} \in \mathcal{C}^{N_t\times 1}$ is the zero-forcing precoding vector, with the assumption of a single layer transmission per user, where $\mathbf{v}^{\kappa}_{k,c} = \left(\mathbf{H}^{\kappa}_{k,c}\right)^{\mathrm{H}}\left(\mathbf{H}^{\kappa}_{k,c}\left(\mathbf{H}^{\kappa}_{k,c}\right)^{\mathrm{H}}\right)^{-1}$. $s^{\kappa}_{k,c}$ and $\mathbf{n}^{\kappa}_{k,c}$ are the transmitted symbol and the additive white Gaussian noise at the k^{th} user, respectively. The first summation indicates the intra-cell interference while the second presents the inter-cell interference, resulted from either the URLLC or eMBB traffic. The 3GPP 3D spatial channel model [31] is adopted, where the DL channel spatial coefficient seen by the m^{th} receive antenna from the n^{th} transmit antenna is composed from Q spatial paths, each with Z rays, and is expressed by

$$h^{\kappa}_{(m,n)_k} = \frac{1}{\sqrt{Q}}\sum_{q=0}^{Q-1}\sqrt{\delta_k}\,\mathcal{G}_{q,k}\,r_{(m,n,q)_k}, \quad (3)$$

where $\delta_k = \ell\epsilon_k^{\varrho}\mu_k$ is a constant, ℓ and μ_k are the propagation and shadow fading factors, respectively, ϵ_k^{ϱ} is the physical distance between transceivers, with ϱ as the pathloss exponent, $\mathcal{G}_{q,k} \sim \mathbb{CN}(0,1)$ is a randomness source per channel path. Hence, the channel steering coefficient $r_{(m,n,q)_k}$ is calculated as

$$r_{(m,n,q)_k} = \sqrt{\frac{\xi\psi}{Z}}\sum_{z=0}^{Z-1}\begin{pmatrix}\sqrt{\mathfrak{D}^{m,n,q,z}_{\mathrm{BS}}(\theta_{\mathrm{AoD}},\varphi_{\mathrm{EoD}})}\,e^{j(\eta d\bar{f}+\Phi_{m,n,q,z})}\\ \sqrt{\mathfrak{D}^{m,n,q,z}_{\mathrm{UE}}(\theta_{\mathrm{AoA}},\varphi_{\mathrm{EoA}})}\,e^{j(\eta d\,\sin(\theta_{\mathrm{m,n,q,z,AoA}}))}\\ e^{j\eta||s||\cos(\varphi_{\mathrm{m,n,q,z,EoA}})\cos(\theta_{\mathrm{m,n,q,z,AoA}}-\theta_s)t}\end{pmatrix}, \quad (4)$$

where ξ and ψ are the power and large-scale coefficients, $\mathfrak{D}_{\mathrm{BS}}$ and $\mathfrak{D}_{\mathrm{UE}}$ are the antenna patterns at the base-station (BS) and UE, respectively, η is the wave number, θ denotes the horizontal angle of arrival θ_{AoA} and departure θ_{AoD}, while φ implies the elevation angle of arrival φ_{EoA} and departure φ_{EoD}, respectively. s is the user speed, $\bar{f} = f_x\cos\theta_{\mathrm{AoD}}\cos\varphi_{\mathrm{EoD}} + f_y\cos\varphi_{\mathrm{EoD}}\sin\theta_{\mathrm{AoD}} + f_z\sin\varphi_{\mathrm{EoD}}$ is the displacement vector of the transmit antenna array (for a uniform linear array, $f_y = f_z = 0$). Accordingly, the received signal at the k^{th} user is decoded by applying the receiver vector $\mathbf{u}^{\kappa}_{k,c}$, given by

$$\left(\mathrm{y}^{\kappa}_{k,c}\right)^* = \left(\mathbf{u}^{\kappa}_{k,c}\right)^{\mathrm{H}}\mathrm{y}^{\kappa}_{k,c}, \quad (5)$$

where $\mathbf{u}^{\kappa}_{k,c}$ is the antenna combining vector, designed by the linear minimum mean square error interference rejection combining (LMMSE-IRC) receiver [32]. Hence, the received SINR at the k^{th} user, assuming an error-free link adaptation process, is expressed by

$$\Upsilon^{\kappa}_{k,c} = \frac{p^c_k\left\|\mathbf{H}^{\kappa}_{k,c}\mathbf{v}^{\kappa}_{k,c}\right\|^2}{1+\sum_{g\in\mathbb{G}_c,g\neq k}p^c_g\left\|\mathbf{H}^{\kappa}_{k,c}\mathbf{v}^{\kappa}_{g,c}\right\|^2+\sum_{j\in C,j\neq c}\sum_{g\in\mathbb{G}_j}p^j_g\left\|\mathbf{H}_{g,j}\mathbf{v}^{\kappa}_{g,j}\right\|^2}, \quad (6)$$

where p^c_k is the k^{th} user receive power. Then, the received per-PRB data rate of the k^{th} user is expressed as

$$r^{\kappa}_{k,rb} = \log_2\left(1+\frac{1}{G_c}\Upsilon^{\kappa}_{k,c}\right). \quad (7)$$

Finally, the effective exponential SNR mapping [33] is applied to map the received SINR levels across $\mathcal{N}$ allocated sub-carriers into one effective SINR as

$$\left(\Upsilon^{\kappa}_{k,c}\right)^{\mathrm{eff.}} = -\mathcal{O}\ln\left(\frac{1}{\mathcal{N}}\sum_{i=1}^{\mathcal{N}}e^{-\frac{\left(\Upsilon^{\kappa}_{k,c}\right)^i}{\mathcal{O}}}\right), \quad (8)$$

where $\mathcal{O}$ is a calibration parameter.

III. PROBLEM FORMULATION

The 5G-NR system performance should be continuously optimized per user-centric, instead of network-centric basis. However, the individual user utility functions are highly correlated and need to be reliably fulfilled, e.g., eMBB rate maximization and URLLC latency minimization as

$$\forall k_{\mathrm{mbb}} \in \mathcal{K}_{\mathrm{mbb}}: \ \arg\max_{\mathcal{K}_{\mathrm{mbb}}}\sum_{k_{\mathrm{mbb}}=1}^{K_{\mathrm{mbb}}}\sum_{rb\in\Xi^{\mathrm{mbb}}_{k_{\mathrm{mbb}}}}\beta_{k_{\mathrm{mbb}}}r^{\mathrm{mbb}}_{k_{\mathrm{mbb}},rb}, \quad (9)$$

$$\forall k_{\mathrm{llc}} \in \mathcal{K}_{\mathrm{llc}}: \ \arg\min_{\mathcal{K}_{\mathrm{llc}}}(\Psi), \quad \text{s.t.} \ \left\|\mathbf{v}^{\kappa}_k\sqrt{\mathrm{P}}\right\|^2, \quad \Psi \leq 1\,\mathrm{ms}, \quad (10)$$

where $\mathcal{K}_{\mathrm{mbb}}$ and $\mathcal{K}_{\mathrm{llc}}$ represent the active sets of eMBB and URLLC users, respectively, $\Xi^{\mathrm{mbb}}_{k_{\mathrm{mbb}}}$ and $\beta_{k_{\mathrm{mbb}}}$ imply the granted set of PRBs and a priority factor of the k^{th} eMBB user. Ψ is the URLLC target one-way latency, assuming a successful first transmission, which can be given by

$$\Psi = \Lambda_{\mathrm{q}} + \Lambda_{\mathrm{bsp}} + \Lambda_{\mathrm{fa}} + \Lambda_{\mathrm{tx}} + \Lambda_{\mathrm{uep}}, \quad (11)$$

where $\Lambda_{\mathrm{q}}, \Lambda_{\mathrm{bsp}}, \Lambda_{\mathrm{fa}}, \Lambda_{\mathrm{tx}}, \Lambda_{\mathrm{uep}}$ are the queuing, BS processing, frame alignment, transmission, and UE processing delays, respectively. Λ_{fa} is upper-bounded by the short TTI interval while Λ_{bsp} & Λ_{uep} are each bounded by 3-OFDM symbol duration [34], due to the enhanced processing capabilities which come with the 5G-NR. Hence, Λ_{q} and Λ_{tx} become the main obstruction against reaching out the hard URLLC latency budget. Λ_{tx} depends on the URLLC outage SINR as

$$\Lambda_{\mathrm{tx}} = \frac{\mathrm{B}_{\mathrm{llc}}}{\left(\Xi^{\mathrm{llc}}_{k_{\mathrm{llc}}} \log_2\left(1 + \frac{\gamma^{\mathrm{llc}}_{k_{\mathrm{llc}}}}{\digamma}\right)\right)}, \tag{12}$$

where $\digamma$ is an outage SINR gap to represent a non-ideal link adaptation process. The URLLC queuing delay Λ_{q} can be mathematically represented by an arbitrary queuing model. For instance, we adopt the $\mathcal{A}/\mathcal{A}/1$ queuing model from data networks theory [35], where the first $\mathcal{A}$ implies a Poisson packet arrival, second $\mathcal{A}$ denotes exponential service times, and notation '1' represents a single layer URLLC transmission. Thus, the mean queuing delay $\overline{\Lambda}_{\mathrm{q}}$, can be expressed as

$$\overline{\Lambda}_{\mathrm{q}} = \frac{1}{\overline{\Lambda}_{\mathrm{tx}}\,(1-\rho)}, \tag{13}$$

where $\rho = \left(\frac{\lambda}{\overline{\Lambda}_{\mathrm{tx}}}\right)$ is the URLLC traffic intensity, with $\overline{\Lambda}_{\mathrm{tx}}$ as the mean transmission time. Thus, in order to achieve the critical URLLC latency, the transmission and queuing delays should be always minimized to provide further allowance for the HARQ re-transmission delay, if the first transmission is not successful.

Fig. 2 depicts the URLLC transmission delay versus the received SINR level for different URLLC payload sizes $\mathrm{B}_{\mathrm{llc}}$ while Fig. 3 describes the associated URLLC queuing delay. As can be observed, with a larger URLLC payload size, a higher SINR point should be always guaranteed to the URLLC UEs in order to reduce the transmission delay. However, the corresponding queuing delay is shown to significantly depend on the URLLC packet arrival rate, e.g., a larger arrival rate with a degraded mean transmission time results in immensely higher queuing delays. This requires allocating

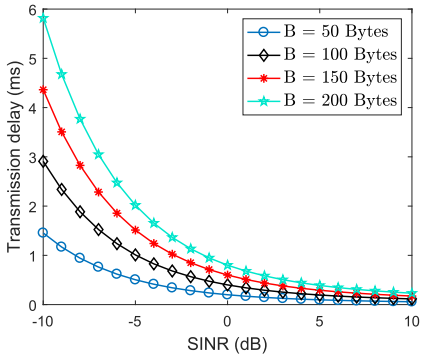

FIGURE 2. URLLC transmission delay with $\mathrm{B}_{\mathrm{llc}}$, $\Xi^{\mathrm{llc}}_{\boldsymbol{k}} = 10$ MHz.

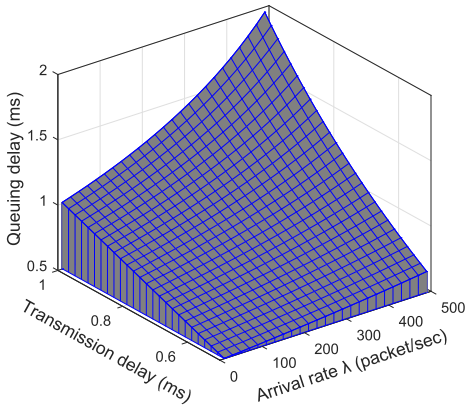

FIGURE 3. URLLC queuing delay with λ and $\overline{\Lambda}_{\mathrm{tx}}$.

excessive radio resources to URLLC traffic or adopting conservative URLLC transmissions. Consequently, the eMBB utility function in (9) is severely under optimized, leading to a significant degradation of the network SE.

IV. PROPOSED SPATIAL PREEMPTIVE SCHEDULING FOR URLLC AND EMBB COEXISTENCE

The proposed NSBPS scheduler seeks to simultaneously cross-optimize the joint performance objectives of the eMBB and URLLC traffic. Thus, the critical URLLC latency deadline is satisfied regardless of the system load while providing the best achievable eMBB performance. When radio resources are not instantly schedulable for incoming URLLC traffic, NSBPS scheduler immediately searches for an ongoing eMBB transmission, that is spatially closest possible to a predefined spatial subspace, i.e., reference subspace. The scheduler instantly projects the selected eMBB transmission onto the reference subspace *on-the-fly*, and accordingly, it assigns the bursty URLLC traffic a portion of the victim eMBB radio resources. At the URLLC user side, it de-orients its decoding vector into one possible null space of the reference subspace; hence, experiencing no inter-user interference, as depicted in Fig. 4. In the following sub-sections, we describe the proposed NSBPS scheduler in-detail.

A. PROPOSED NSBPS – AT THE BS SIDE

Starting at an arbitrary TTI instance, the newly arrived or buffered eMBB traffic is scheduled over single-user (SU) dedicated resources, if there are no pending URLLC arrivals. To dynamically multiplex the active eMBB user allocations across available resources, the proportional fair (PF) scheduling criterion [36] is applied as

$$\Theta_{\mathrm{PF}} = \frac{r^{\mathrm{mbb}}_{k_{\mathrm{mbb}},rb}}{\overline{r}^{\mathrm{mbb}}_{k_{\mathrm{mbb}},rb}}, \tag{14}$$

$$k^{*}_{\mathrm{mbb}} = \arg\max_{\mathcal{K}_{\mathrm{mbb}}} \Theta_{\mathrm{PF}}, \tag{15}$$

where $\overline{r}^{\mathrm{mbb}}_{k_{\mathrm{mbb}},rb}$ is the average delivered data rate of the k^{th} eMBB user. However, in case of URLLC new DL arrivals at the BS while sufficient schedulable resources are instantly

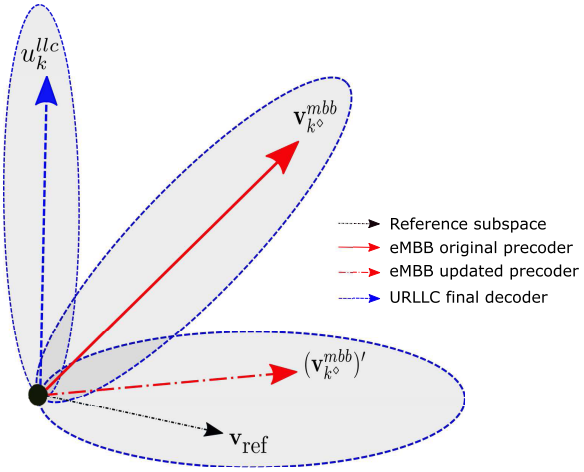

FIGURE 4. NSBPS scheduler: eMBB precoder projection and URLLC decoder orientation.

available, the NSBPS scheduler overwrites the eMBB user SU scheduling priority for the sake of the newly arrived URLLC traffic, by the weighted PF scheduling criteria (WPF) as

$$\Theta_{\text{WPF}} = \frac{r^{\kappa}_{k_\kappa,rb}}{\overline{r}^{\kappa}_{k_\kappa,rb}}\beta_{k_\kappa}, \tag{16}$$

with $\beta_{k_{\text{llc}}} \gg \beta_{k_{\text{mbb}}}$ for immediate URLLC SU scheduling.

Nonetheless, with a large offered loading level, which is foreseen with the 5G-NR, sufficient resource allocation may not be instantly available for the incoming URLLC traffic. For example, URLLC packets may arrive at the BS during an eMBB transmission slot (14-OFDM symbol). Hence, larger scheduling delays, i.e., queuing and/or segmentation delays, are experienced. The URLLC segmentation delay indicates that arrived URLLC payload is segmented and transmitted over multiple TTIs, due to insufficient instant resource allocation or degraded capacity per PRB. For such case, the proposed NSBPS scheduler first attempts fitting the URLLC traffic within one active eMBB transmission using a standard and non-biased MU-MIMO transmission, and based on a highly conservative $\gamma-$orthogonality threshold, with $\gamma \rightarrow [0, 1]$. Thus, incoming URLLC traffic can only be paired with an active eMBB transmission if:

$$1 - \left|\left(\mathbf{v}^{\text{mbb}}_{k_{\text{mbb}}}\right)^{\text{H}} \mathbf{v}^{\text{llc}}_{k_{\text{llc}}}\right|^2 \geq \gamma. \tag{17}$$

The conservative, i.e., large, orthogonality threshold is forcibly applied to protect the URLLC traffic against potential inter-user interference. If the system spatial degrees of freedom (SDoFs) are restrained during an arbitrary TTI and such large orthogonality requirements can not be satisfied, the NSBPS scheduler instantly enforces a semi-transparent, i.e., URLLC-aware transmission, controlled, i.e., independently from the available SDoFs, and biased, i.e., for the sake of URLLC user end, MU-MIMO transmission. The URLLC outage requirements are then achieved by satisfying:

$$\textbf{rank}\left\{\left(\mathbf{u}^{\text{llc}}_k\right)^{\text{H}} \mathbf{H}^{\text{llc}}_k \mathbf{v}^{\text{llc}}_k\right\} \sim \text{full}, \tag{18}$$

$$\textbf{rank}\left\{\left(\mathbf{u}^{\text{llc}}_k\right)^{\text{H}} \mathbf{H}^{\text{llc}}_k \left(\mathbf{v}^{\text{mbb}}_{k^\diamond}\right)'\right\} \sim 0, \tag{19}$$

where $\left(\mathbf{v}^{\text{mbb}}_{k^\diamond}\right)'$ is the actual precoder of the co-scheduled eMBB user with the incoming URLLC user. Then, an arbitrary spatial subspace is pre-defined in the discrete Fourier transform beamforming domain [37] as

$$\mathbf{v}_{\text{ref}}(\theta) = \left(\frac{1}{\sqrt{N_t}}\right)\left[1, e^{-j2\pi\,\Delta\cos\theta}, \ldots, e^{-j2\pi\,\Delta(N_t-1)\cos\theta}\right]^{\text{T}}, \tag{20}$$

where Δ is the absolute antenna spacing and θ is an arbitrary spatial angle. Accordingly, the NSBPS scheduler searches for one active eMBB user whose transmission is most aligned within the reference subspace $\mathbf{v}_{\text{ref}}(\theta)$ as

$$k^{\diamond}_{\text{mbb}} = \arg\min_{\mathcal{K}_{\text{mbb}}} \mathbf{d}\left(\mathbf{v}^{\text{mbb}}_k, \mathbf{v}_{\text{ref}}\right), \tag{21}$$

where the Chordal distance $\mathbf{d}$ between $\mathbf{v}^{\text{mbb}}_k$ and $\mathbf{v}_{\text{ref}}$ is expressed by

$$\mathbf{d}\left(\mathbf{v}^{\text{mbb}}_k, \mathbf{v}_{\text{ref}}\right) = \frac{1}{\sqrt{2}}\left\|\mathbf{v}^{\text{mbb}}_k\left(\mathbf{v}^{\text{mbb}}_k\right)^{\text{H}} - \mathbf{v}_{\text{ref}}\mathbf{v}^{\text{H}}_{\text{ref}}\right\|. \tag{22}$$

Next, the NSBPS scheduler applies an instant precoder projection of the selected victim eMBB user $\mathbf{v}^{\text{mbb}}_{k^\diamond}$ onto $\mathbf{v}_{\text{ref}}$ as given by

$$\left(\mathbf{v}^{\text{mbb}}_{k^\diamond}\right)' = \frac{\mathbf{v}^{\text{mbb}}_{k^\diamond}\cdot\,\mathbf{v}_{\text{ref}}}{\|\mathbf{v}_{\text{ref}}\|^2} \times \mathbf{v}_{\text{ref}}, \tag{23}$$

wherein $\left(\mathbf{v}^{\text{mbb}}_{k^\diamond}\right)'$ is the post-projection updated eMBB user precoder. This way, the NSBPS scheduler immediately schedules the sporadic URLLC traffic over partial or full shared resource allocation with the victim eMBB transmission. Thus, in principal, no URLLC queuing delays are experienced. On another side, due to the instant projection of the victim eMBB user precoder, it exhibits a capacity loss; however, it is highly constrained and only limited by the spatial projection loss over the shared resources with the URLLC traffic. Furthermore, under larger eMBB user loading, the NSBPS scheduler is highly likely to find an active eMBB user whose transmission is originally aligned within the reference spatial subspace; hence, the instant spatial projection would not significantly impact its achievable capacity. Finally, the BS transmits a single-bit co-scheduling true indication, i.e., $\alpha = 1$, to the intended URLLC user, which is transmitted in the user-centric CCH.

B. PROPOSED NSBPS – AT THE URLLC USER SIDE

When a true co-scheduling indication $\alpha = 1$ is detected, the URLLC user acknowledges that its resource allocation is shared with an active eMBB transmission, whose interference is limited within the reference subspace.

Thus, the URLLC user first designs its decoder vector using a standard LMMSE-IRC receiver, to reject inter-cell interference as

$$\left(\mathbf{u}_k^{\text{llc}}\right)^{(1)} = \left(\mathbf{H}_k^{\text{llc}}\mathbf{v}_k^{\text{llc}}\left(\mathbf{H}_k^{\text{llc}}\mathbf{v}_k^{\text{llc}}\right)^{\text{H}} + \mathbf{W}\right)^{-1}\mathbf{H}_k^{\text{llc}}\mathbf{v}_k^{\text{llc}}, \quad (24)$$

where the interference covariance matrix is given by

$$\mathbf{W} = \mathbb{E}\left(\mathbf{H}_k^{\text{llc}}\mathbf{v}_k^{\text{llc}}\left(\mathbf{H}_k^{\text{llc}}\mathbf{v}_k^{\text{llc}}\right)^{\text{H}}\right) + \sigma^2\text{I}_{M_r}, \quad (25)$$

where I_{M_r} is $M_r \times M_r$ identity matrix. The decoder vector statistics $\left(\mathbf{u}_k^{\text{llc}}\right)^{(1)}$ are then transferred to one possible null space of the observed effective inter-user interference subspace $\mathbf{H}_k^{\text{llc}}\mathbf{v}_{\text{ref}}$, as given by

$$\left(\mathbf{u}_k^{\text{llc}}\right)^{(2)} = \left(\mathbf{u}_k^{\text{llc}}\right)^{(1)} - \frac{\left(\left(\mathbf{u}_k^{\text{llc}}\right)^{(1)} \cdot \mathbf{H}_k^{\text{llc}}\mathbf{v}_{\text{ref}}\right)}{\left\|\mathbf{H}_k^{\text{llc}}\mathbf{v}_{\text{ref}}\right\|^2} \times \mathbf{H}_k^{\text{llc}}\mathbf{v}_{\text{ref}}. \quad (26)$$

Accordingly, the final URLLC decoder vector $\left(\mathbf{u}_k^{\text{llc}}\right)^{(2)}$ experiences no inter-user interference, providing the URLLC user with a robust decoding ability. To summarize the major concept of the proposed NSBPS scheduler, Fig. 5 shows a high level flow diagram of the NSBPS scheduler at the BS and intended URLLC user, respectively.

V. ANALYTICAL ANALYSIS COMPARED TO STATE OF THE ART URLLC SCHEDULERS

In this section, we introduce an analytical performance comparison of the proposed NSBPS scheduler versus the state-of-the-art schedulers from industry and academia as follows:

1. Punctured scheduler (PS) [22]: in case that sufficient radio resources are not instantly available for the sporadic URLLC traffic, the PS scheduler immediately overwrites part of the ongoing eMBB transmissions by the incoming URLLC traffic. Thus, in principal, the URLLC queuing delay component is significantly minimized. PS scheduler has shown sound improvement of the URLLC latency performance; however, with a highly degraded SE, due to the eMBB unrealizable punctured transmissions.

2. Enhanced punctured scheduler (E-PS) [23]: E-PS scheduler is an improved version of the conventional PS scheduler, which is recently proposed to partially recover the lost eMBB capacity due to puncturing. Punctured eMBB UEs are presumed to be aware of which resources are being punctured by URLLC traffic. Thus, victim eMBB UEs disregard the punctured PRBs from the Chase combining HARQ process in order not to spread the decoding errors. Furthermore, two code-block (CB) mapping layouts [23]–[25] are evaluated as: fully interleaved (FI), and frequency first (FF) layouts, respectively. The former indicates that CBs associated with an eMBB transport block (TB) are fully interleaved over the time and frequency resources, however, the latter means that CBs are spread over the frequency domain and condensed over the time domain. Moreover, CB-based HARQ feedback is adopted in order for the impacted eMBB UEs to feedback the BS of which punctured CBs could not be successfully decoded, hence, only re-transmitting the victim CBs instead of the full TB, reducing the aggregate HARQ overhead.

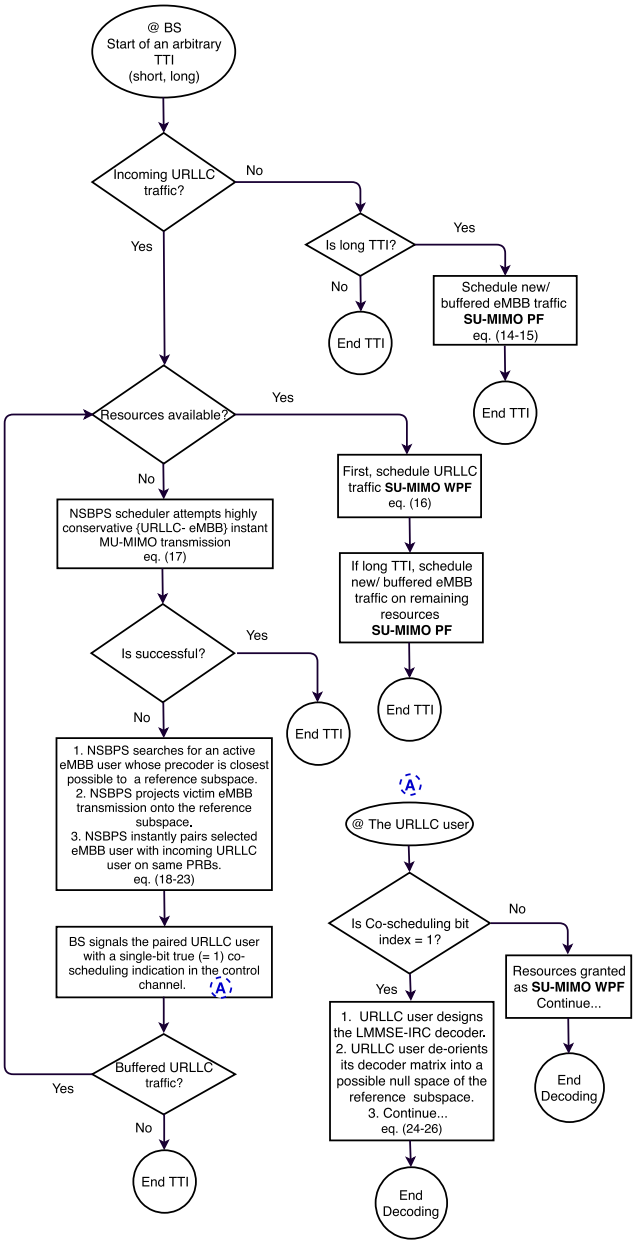

FIGURE 5. Flow diagram of the NSBPS scheduler, at the BS and the intended URLLC user, respectively.

3. Multi-user punctured scheduler (MU-PS) [26]: in our recent work, we considered a MU transmission on top of the PS scheduler. The proposed MU-PS scheduler first attempts a non-biased and transparent MU transmission of an URLLC-eMBB user pair. If the system offered SDoFs during an arbitrary TTI are not sufficient, the MU-PS scheduler rolls back to PS scheduler, where the URLLC traffic immediately punctures part of the radio resources, monopolized by ongoing eMBB transmissions. The MU-PS exhibits a fair tradeoff between URLLC latency and overall SE. However, the achievable MU gain is shown to be

very restrained with the SDoF-limited conditions, where the MU-PS scheduler is highly likely to fall back to PS scheduler. Furthermore, it has been demonstrated that MU-PS scheduler leads to a degradation of the URLLC decoding ability, due to the potential inter-user interference. Thus, a conservative MU-PS (CMU-PS) scheduler is introduced to further safe-guard the URLLC traffic against potentially strong inter-user interference, even if the pairing sum capacity constraint is satisfied. Thus, users can only be paired in a MU-MIMO transmission if their precoders satisfy larger spatial separation as given by

$$\left| \angle \left(\mathbf{v}^{\text{mbb}}_{k_{\text{mbb}}} \right) - \angle \left(\mathbf{v}^{\text{llc}}_{k_{\text{llc}}} \right) \right| \geq \vartheta, \tag{27}$$

where ϑ is a predefined spatial separation threshold.

Accordingly, the aggregate eMBB user rate is calculated from the individual sub-carrier rates, assuming OFDMA flat fading channels, as

$$r^{\text{mbb}}_{k_{\text{mbb}}} = \Xi^{\text{mbb}}_{k_{\text{mbb}}} r^{\text{mbb}}_{k_{\text{mbb}},rb}. \tag{28}$$

Next, the fraction of the resources $\Gamma^{\text{llc}}_{k_{\text{mbb}}}$, allocated to the k^{th} eMBB user and being altered by the incoming URLLC traffic, is expressed as a set of random variables, given by

$$\mathbf{\Gamma} = \left(\Gamma^{\text{llc}}_{k_{\text{mbb}}} \mid k_{\text{mbb}} \in \mathcal{K}_{\text{mbb}} \right). \tag{29}$$

Due to the small size of URLLC packets, it is reasonable to assume that $\Gamma^{\text{llc}}_{k_{\text{mbb}}} \leq \Xi^{\text{mbb}}_{k_{\text{mbb}}}$ is satisfied. The achievable eMBB user rate can then be formulated by the joint eMBB and URLLC rate allocation function, as expressed by

$$R_{k_{\text{mbb}}} = \mathcal{F} \left(\Xi^{\text{mbb}}_{k_{\text{mbb}}}, \Gamma^{\text{llc}}_{k_{\text{mbb}}} \right). \tag{30}$$

For example, an eMBB user exhibits no capacity loss if its associated resource allocation is not induced by incoming URLLC traffic, hence, $\mathcal{F} \left(\Xi^{\text{mbb}}_{k_{\text{mbb}}}, \Gamma^{\text{llc}}_{k_{\text{mbb}}} \right) = \Xi^{\text{mbb}}_{k_{\text{mbb}}} r^{\text{mbb}}_{k_{\text{mbb}},rb}$. However, since the URLLC traffic is always prioritized, victim eMBB users exhibit a rate loss over a fraction of the impacted PRBs, where it can be formulated by the rate loss function Π as

$$\mathcal{F} \left(\Xi^{\text{mbb}}_{k_{\text{mbb}}}, \Gamma^{\text{llc}}_{k_{\text{mbb}}} \right) = \Xi^{\text{mbb}}_{k_{\text{mbb}}} r^{\text{mbb}}_{k_{\text{mbb}},rb} \left(1 - \Pi \right), \tag{31}$$

where the rate loss function $\Pi : [0, 1] \rightarrow [0, 1]$ represents the effective portion of impacted PRBs of the k^{th} eMBB user. Under the proposed NSBPS framework, the updated eMBB effective channel gain is expressed as

$$\mathcal{Q}^{\text{mbb}}_{k} = \frac{1}{\left[\left(\mathbf{H}^{\text{mbb}}_{k} \left(\mathbf{v}^{\text{mbb}}_{k^{\diamond}} \right)' \right) \times \left(\mathbf{H}^{\text{mbb}}_{k} \left(\mathbf{v}^{\text{mbb}}_{k^{\diamond}} \right)' \right)^{\text{H}} \right]^{-1}}, \tag{32}$$

where $\mathcal{Q}^{\text{mbb}}_{k}$ is the post-projection channel gain of the k^{th} eMBB user. The magnitude of $\mathcal{Q}^{\text{mbb}}_{k}$ can be reformulated in terms of the eMBB projection loss, due to the immediate change of the eMBB precoder from $\mathbf{v}^{\text{mbb}}_{k^{\diamond}}$ to $\left(\mathbf{v}^{\text{mbb}}_{k^{\diamond}} \right)'$, as

$$\mathcal{Q}^{\text{mbb}}_{k} = \left\| \mathbf{H}^{\text{mbb}}_{k} \mathbf{v}^{\text{mbb}}_{k^{\diamond}} \right\|^2 \times \sin^2 \left(\theta_{\left[\mathbf{v}^{\text{mbb}}_{k^{\diamond}}, \left(\mathbf{v}^{\text{mbb}}_{k^{\diamond}} \right)' \right]} \right), \tag{33}$$

where $\sin^2 \left(\theta_{\left[\mathbf{v}^{\text{mbb}}_{k^{\diamond}}, \left(\mathbf{v}^{\text{mbb}}_{k^{\diamond}} \right)' \right]} \right)$ denotes the eMBB precoder projection loss, over the shared resources with the URLLC traffic, and $\theta_{\left[\mathbf{v}^{\text{mbb}}_{k^{\diamond}}, \left(\mathbf{v}^{\text{mbb}}_{k^{\diamond}} \right)' \right]}$ is the spatial angle discrepancy between its original and projected precoders, respectively. Thus, $\overset{\text{NSBPS}}{\Pi}$ is estimated as

$$\overset{\text{NSBPS}}{\Pi} = \left(\frac{\Gamma^{\text{llc}}_{k_{\text{mbb}}}}{\Xi^{\text{mbb}}_{k_{\text{mbb}}}} \right) \times \sin^2 \left(\theta_{\left[\mathbf{v}^{\text{mbb}}_{k^{\diamond}}, \left(\mathbf{v}^{\text{mbb}}_{k^{\diamond}} \right)' \right]} \right). \tag{34}$$

Due to the constraints in (17) and (21), the eMBB projection loss is guaranteed minimum at all times since:

$$\sin^2 \left(\theta_{\left[\mathbf{v}^{\text{mbb}}_{k^{\diamond}}, \left(\mathbf{v}^{\text{mbb}}_{k^{\diamond}} \right)' \right]} \right) \ll 1. \tag{35}$$

On another side, the rate loss function of the PS scheduler is expressed by the full URLLC resources altering the eMBB user resources, since the eMBB transmission is instantly stopped over these resources, and it is given by

$$\overset{\text{PS}}{\Pi} = \left(\frac{\Gamma^{\text{llc}}_{k_{\text{mbb}}}}{\Xi^{\text{mbb}}_{k_{\text{mbb}}}} \right). \tag{36}$$

The MU-PS scheduler provides an optimized average of the achievable eMBB user rate; however, the MU gain is constrained by the available SDoFs, due to the persistent PS events, if the standard MU-MIMO scheduler fails. Hence, the MU-PS rate loss can be given by

$$\overset{\text{MU-PS}}{\Pi} = \phi \left(\frac{\Gamma^{\text{llc}}_{k_{\text{mbb}}}}{\Xi^{\text{mbb}}_{k_{\text{mbb}}}} \right), \tag{37}$$

where $\phi \leq 1$ is the probability of rolling back to PS scheduler under a given cell loading state. Fig. 6 presents the discrete values of ϕ under different loading conditions, where we define the cell loading as: $\Omega = (K_{\text{mbb}}, K_{\text{llc}})$, and the eMBB full buffer traffic is adopted. As can be observed, with a small number of eMBB users per cell, the system overall SDoFs are highly limited and hence, the MU-PS scheduler is highly likely to roll back to PS scheduler, i.e., $\phi \sim 1$

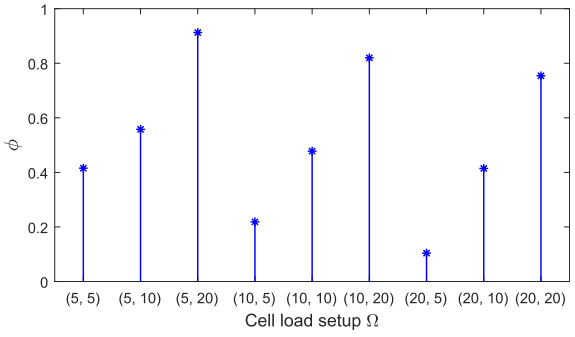

FIGURE 6. MU-PS scheduler: discrete probabilities ϕ of falling back to PS scheduler with Ω.

TABLE 1. Simulation setup and major parameters.

Parameter	Value
Network environment	3GPP Urban Macro (UMa) network with 21 cells and 500 meter inter-site distance
Carrier configuration	10 MHz carrier bandwidth at 2 GHz
Propagation	$128.1 + 37.6\log(D[km])$ dB; Log-Normal shadowing with 8 dB standard deviation
PHY numerology	15 kHz subcarrier spacing; 12 subcarriers per PRB; 2-OFDM symbols TTI (0.143 ms)
Control channel	Error-free in-resource scheduling grants with dynamic link adaptation
Data channel MCS	QPSK to 64QAM, LTE coding rates
CSI	LTE-alike CQI and PMI, $\tilde{a} = 0.01$, reported every 5 ms; Sub-band size: 8 PRBs
Antenna configuration	8 x 2 MU-MIMO with LTE-alike precoding and LMMSE-IRC receiver
Packet scheduler	Weighted Proportional Fair with priority for URLLC traffic
BLER target	eMBB: 10 percent; URLLC: 1 percent
HARQ	Asynchronous HARQ with Chase combining and 4 TTI round trip time; Max. 6 HARQ retransmissions
RLC setup	Transparent mode
Traffic composition	URLLC: 5, 10, and 20 users / cell; eMBB: 5, 10, and 20 users / cell
UE distribution	Uniformly distributed; 3 km/h UE speed
Traffic model	URLLC: FTP3 downlink traffic, $\lambda = 250$ and B = 50 bytes; eMBB: full buffer and CBR traffic of 4 Mbps load per user

in order to instantly schedule the offered URLLC traffic. Finally, the average achievable eMBB user rate is expressed by

$$\overline{R}_{k_{\text{mbb}}} = \Xi^{\text{mbb}}_{k_{\text{mbb}}} r^{\text{mbb}}_{k_{\text{mbb}},rb} \left(1 - \mathbb{E}\left(\Pi\right)\right). \tag{38}$$

Based on (28) - (38), it can be further observed that the proposed NSBPS scheduler exhibits the highest ergodic capacity, due to the constrained eMBB rate loss function.

VI. PERFORMANCE EVALUATION

The performance of the NSBPS scheduler is validated by extensive SLSs, where the major 5G-NR and radio resource management functionalities are implemented, e.g., agile frame structure, HARQ re-transmission, dynamic link adaptation, and control channel overhead, as described in the subsection II-A. The major simulation parameters are listed in Table 1. The baseline antenna configuration is 8×2 and the default eMBB traffic is full buffer unless otherwise mentioned.

A. MAJOR PERFORMANCE COMPARISON

Fig. 7 depicts the one-way latency of the URLLC traffic at the 10^{-5} outage probability under different cell loading conditions Ω, for the proposed NSBPS, PS, MU-PS, and time-domain WPF (TD-WPF) schedulers. As can be noticed, the NSBPS scheduler offers significant robustness of the URLLC latency performance, independently from the cell loading conditions, and hence, the aggregate interference levels. The performance gain of the NSBPS scheduler is attributed to: a) the elimination of the scheduling queuing delays of the URLLC sporadic traffic, i.e., guaranteed instant URLLC scheduling, b) safeguarding the URLLC traffic from the potential inter-user interference through controlled (almost surely occurs), biased (in favor of the URLLC user), and semi-transparent MU-MIMO transmission, c) compressing the interference spatial dimension, leading to a better

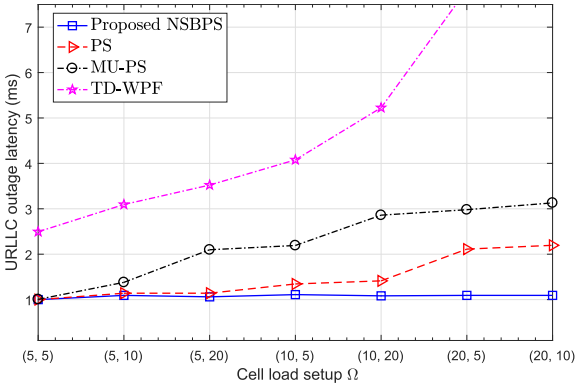

FIGURE 7. URLLC outage latency of the NSBPS, PS, MU-PS, and TD-WPF schedulers,with Ω.

LMMSE receiver interference rejection ability, as will be presented in subsection VI-B, and d) the always constrained minimum eMBB cost function.

The PS scheduler provides an optimized URLLC latency performance, especially over the low load region; however, it comes at the expense of a degraded SE. Moreover, it exhibits URLLC performance degradation as the cell load increases, due to the resulting extreme levels of inter-cell interference. Accordingly, a degraded capacity per PRB is experienced. The MU-PS scheduler provides a decent trade-off between URLLC latency and overall SE due to the achievable MU gain. However, the non-controlled MU interference degrades the URLLC decoding point, especially when the inter-cell interference levels are originally significant. Finally, the TD-WPF scheduler exhibits the worst latency performance since instant URLLC scheduling is not guaranteed, e.g., the URLLC packets are queued for multiple TTIs if the instant schedulable radio resources are not sufficient to accommodate these payloads.

Fig. 8 shows the average cell throughput in Mbps with the cell loading condition Ω. The NSBPS scheduler achieves

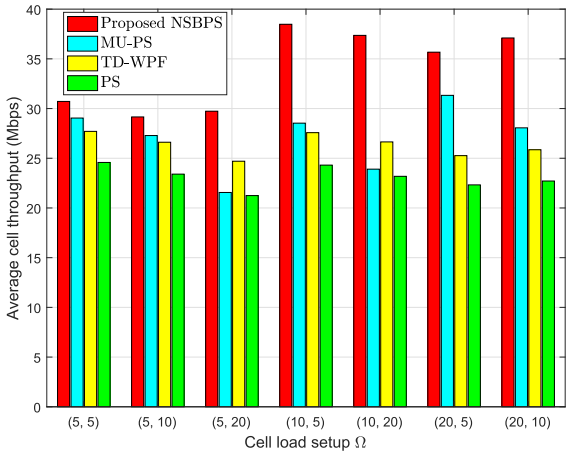

FIGURE 8. Average cell throughput performance of the NSBPS, PS, MU-PS, and TD-WPF schedulers, with Ω.

the best cell throughput performance because the eMBB cost function is limited by the spatial projection loss, and thus, it is always constrained minimum, compared to the PS, MU-PS, and TD-WPF schedulers. The PS scheduler clearly suffers from severe degradation in the cell ergodic capacity due to the eMBB punctured transmissions. However, the TD-WPF scheduler exhibits an improved cell performance since punctured eMBB transmissions are not allowed; however, at the expense of significant URLLC queuing delays. Finally, the MU-PS scheduler provides a better cell capacity than TD-WPF and PS schedulers, due to the achieved MU gain; however, gain is highly limited by the available system SDoFs, and hence, dependent on the cell loading condition, and aggregate interference levels, e.g., MU-PS scheduler is highly likely to roll back to SE-less-efficient PS scheduler when the system SDoFs are limited within a TTI. In Fig. 9, we compare the empirical cumulative distribution function (ECDF) of the achievable cell throughput of the proposed NSBPS scheduler against the state-of-the-art E-PS and CMU-PS schedulers, respectively, for $\Omega = (5, 5)$. As can be clearly identified, the NSBPS scheduler still outperforms all schedulers under assessment due to the guaranteed minimum projection loss of the victim eMBB UEs. On the other hand, the CMU-PS scheduler provides an optimized cell throughput performance due to enforcing a conservative MU pairing constraint; thus, the CMU-PS scheduler performs less MU pairings; however, with a higher MU gain. The conventional PS scheduler shows the worst SE because the puncturing events severely degrade the eMBB capacity. Finally, the E-PS scheduler shows an improved cell throughput than the PS, for both the FI and FF CB layouts, respectively. The E-PS scheduler with FI CB layout is shown to slightly outperform that is of the FF CB, since a modest and equal puncturing impact on all CBs minimizes the error probability of the entire TB compared to the case of the FF CB, where only a few CBs, i.e., condensed in the time-domain, are completely damaged due to puncturing.

B. PERFORMANCE DRIVERS OF THE PROPOSED NSBPS SCHEDULER

Examining the performance drivers of the proposed NSBPS scheduler, Fig. 10 shows the average achievable capacity per scheduled eMBB/URLLC allocations in bits. The proposed scheduler clearly enhances the allocation average capacity due to the controlled MU pairing, and the limited eMBB projection loss. The MU-PS scheduler shows an improved capacity, however, it depends on the available system SDoFs, e.g., with SDoF-limited condition ($\Omega = (5, 20)$), the MU-PS scheduler exhibits a similar allocation capacity as of the PS scheduler. The PS scheduler provides the worst performance due to the punctured eMBB transmissions and the hard priority of the URLLC traffic. Similar conclusions can be also reached from Fig. 11, where the average number of the TD queued users is depicted, i.e., the average number of active users which are queued in the TD scheduler for multiple TTIs until sufficient resources are released. Due to its achievable higher allocation capacity, the NSBPS scheduler shows the lowest number of the TD-queued users against the MU-PS and PS schedulers, respectively. However, under a large offered load, e.g., $\Omega = (20, 5)$, all schedulers under evaluation suffer from a larger queuing delay due to the extreme interference levels, and hence, PRB degraded capacity. Furthermore, Fig. 12 depicts the URLLC per packet effective SINR in dB, as in eq. (8), for $\Omega = (5, 20)$. The NSBPS

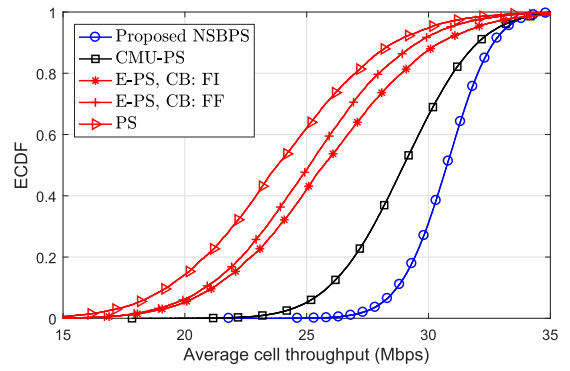

FIGURE 9. Average cell throughput performance of the NSBPS, CMU-PS, E-PS, and PS schedulers, with $\Omega = (5, 5)$.

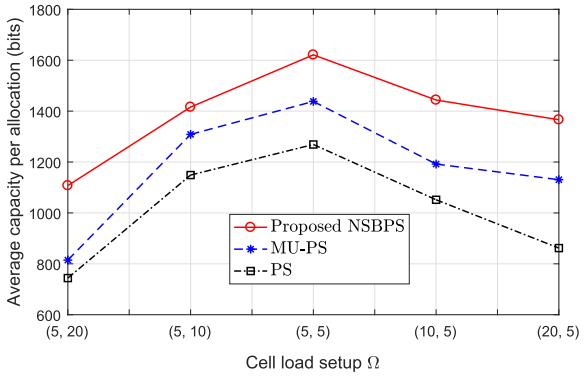

FIGURE 10. Average capacity per scheduled allocation size of the NSBPS, MU-PS, and PS schedulers, with Ω.

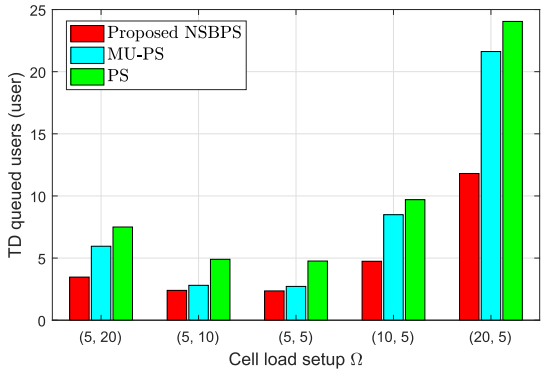

FIGURE 11. TD user queuing performance of the NSBPS, MU-PS, and PS schedulers, with Ω.

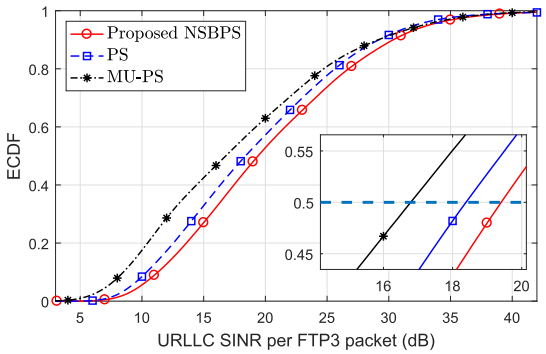

FIGURE 12. URLLC per packet SINR performance of the NSBPS, MU-PS, and PS schedulers, with $\Omega = (5, 20)$.

scheduler provides $\sim$ 1 dB gain in the average FTP3 packet SINR over the PS scheduler. The fixed subspace projection of the victim eMBB transmissions leads to regularizing the inter-cell interference statistics from different cells into a compressed spatial span. Thus, the LMMSE-IRC receiver has better SDoFs to reject and null the interference statistics from the received signal, leading to a better SINR performance with the NSBPS scheduler. However, the MU-PS scheduler exhibits the worst SINR level per FTP3 packet due to the residual inter-user interference from the standard MU transmissions.

C. EMBB REALISTIC TRAFFIC MODEL

Examining the end-to-end eMBB performance, we also consider a more realistic traffic modeling in order to emulate the coexistence of the broadband video streaming services with the URLLC applications. Under this assumption, a constant bit rate (CBR) traffic modeling is adopted for the eMBB users, where $\check{n} = 10$, $B_{mbb} = 320$ KBytes, and $\check{\imath} = 0.6864$ sec. This implies a clip time of $\sim$ 6.1776 sec and CBR load of $\sim$ 4 Mbps per eMBB user. When an arbitrary eMBB user finishes its corresponding streaming session, another eMBB user is generated with a random position in the simulation.

Fig. 13 depicts the complementary CDF (CCDF) of the URLLC one-way latency, for different antenna configurations, i.e., 8 $\times$ 2 and 8 $\times$ 8, respectively. As can

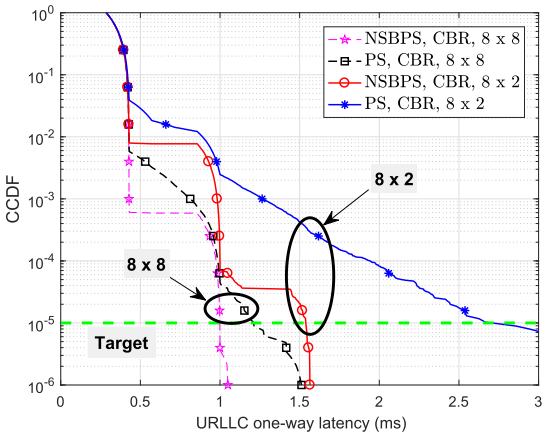

FIGURE 13. URLLC latency CCDF of the NSBPS, and PS schedulers, with eMBB CBR traffic and $\Omega = (5, 10)$.

be seen, with 8 $\times$ 2 antenna setup, the URLLC latency performance of both NSBPS and PS schedulers is significantly degraded, where the URLLC 1 ms outage latency can not be satisfied. This is due to the highly varying set of active interferers, resulting from the bursty eMBB CBR traffic. Hence, the resultant fast varying interference pattern disrupts the URLLC link adaptation process, leading to several HARQ re-transmissions before a successful decoding. One possible suggestion is to utilize the channel hardening phenomenon [38] by increasing the size of the transmit and receive antenna arrays, for the same transceiver complexity. With larger antenna arrays, the spatial channel becomes more directive on the desired paths with much less energy leakage on interference paths, leading to a better decoding ability of the LMMSE-IRC receiver. Hence, with 8 $\times$ 8 antenna setup, the URLLC latency performance of both schedulers is clearly improved, achieving the URLLC latency target with the NSBPS scheduler, due to the significantly reduced interference leakage. Finally, Fig. 14 depicts the ECDF of the achievable eMBB user CBR, where similar conclusions can be drawn.

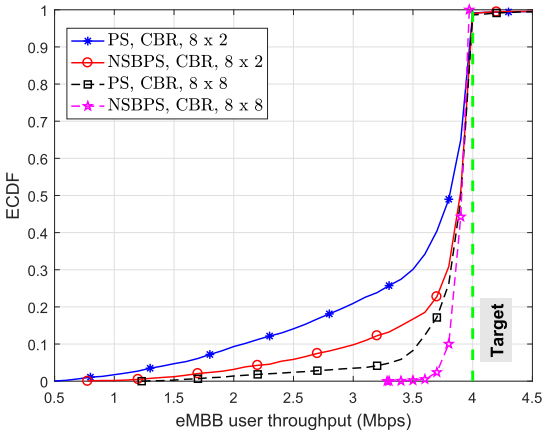

FIGURE 14. eMBB CBR throughput performance of the NSBPS, and PS schedulers, with $\Omega = (5, 10)$.

VII. CONCLUDING REMARKS

An attractive null-space-based preemptive scheduler (NSBPS) for joint eMBB and URLLC traffic is introduced. Proposed NSBPS scheduler guarantees an instant scheduling for the sporadic URLLC traffic, and with the minimal impact on the overall ergodic capacity. Thus, the sporadic URLLC traffic experiences no further queuing delays in order to achieve its critical one-way latency budget. A variety of dynamic system level simulations in addition to an analytic analysis of the major performance indicators are carried out to validate the performance of the proposed scheduler. Compared to the state-of-the-art scheduling proposals from industry and academia, the proposed NSBPS shows extreme URLLC latency robustness with significantly improved eMBB performance.

The major conclusions brought by this paper can be summarized as follows: (1) the transmission and queuing delay components are the major obstacles against achieving the URLLC hard latency, and those are highly correlated and dependent on the URLLC payload size and the mean packet arrival rate, (2) thus, URLLC users must satisfy their outage capacity of interest instead of the overall ergodic capacity, leading to a severe degradation of the network spectral efficiency, (3) proposed NSBPS scheduler instantly schedules the sporadic URLLC traffic regardless of the network loading state, reducing the URLLC queuing delays, and (4) NSBPS scheduler safeguards the URLLC traffic from potential inter-user interference by enforcing sufficient spatial separation through subspace projection. A detailed study on recovering the eMBB capacity will be considered in a future work.

ACKNOWLEDGMENTS

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 (Release 15) V2.0.0*, document TS 38.300,3GPP, Dec. 2017.
[2] *Study on New Radio Access Technology (Release 14) V14.0.0*, document TR 38.801, 3GPP, Mar. 2017.
[3] *Study on Scenarios and Requirements for Next Generation Access Technologies (Release 14) V14.3.0*, document TR 38.913, 3GPP, Jun. 2016.
[4] *Study on New Radio Access Technology Physical Layer Aspects (Release 14) V14.2.0*, document TR 38.802, 3GPP, Sep. 2017.
[5] *Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond*, document ITU-R M.2083-0, International Telecommunication Union (ITU), Feb. 2015.
[6] P. Popovski, "Ultra-reliable communication in 5G wireless systems," in *Proc. IEEE Int. Conf. 5G Ubiquitous Connectivity*, Akaslompolo, Finland, Nov. 2014, pp. 146–151.
[7] E. Dahlman *et al.*, "5G wireless access: Requirements and realization," *IEEE Commun. Mag.*, vol. 52, no. 12, pp. 42–47, Dec. 2014.
[8] M. Simsek, A. Aijaz, M. Dohler, J. Sachs, and G. Fettweis, "5G-enabled tactile Internet," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 3, pp. 460–473, Mar. 2016.
[9] B. Soret, P. Mogensen, K. I. Pedersen, and M. C. Aguayo-Torres, "Fundamental tradeoffs among reliability, latency and throughput in cellular networks," in *Proc. IEEE Globecom*, Austin, TX, USA, Dec. 2014, pp. 1391–1396.
[10] 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.
[11] A. A. Esswie and K. I. Pedersen, "Null space based preemptive scheduling for joint URLLC and eMBB traffic in 5G networks," in *Proc. IEEE Globecom*, Abu Dhabi, United Arab Emirates, Dec. 2018.
[12] 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.
[13] G. Pocovi, K. I. Pedersen, and P. Mogensen, "Joint link adaptation and scheduling for 5G ultra-reliable low-latency communications," *IEEE Netw.*, vol. 6, pp. 28912–28922, May 2018.
[14] K. I. Pedersen, G. Berardinelli, F. Frederiksen, P. Mogensen, and A. Szufarska, "A flexible 5G frame structure design for frequency-division duplex cases," *IEEE Commun. Mag.*, vol. 54, no. 3, pp. 53–59, Mar. 2016.
[15] 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*, Washington, DC, USA, Dec. 2016, pp. 1–7.
[16] G. Pocovi, B. Soret, M. Lauridsen, K. I. Pedersen, and p. Mogensen, "Signal quality outage analysis for ultra-reliable communications in cellular networks," in *Proc. IEEE Globecom*, San Diego, CA, USA, Dec. 2015, pp. 1–6.
[17] G. Pocovi, M. Lauridsen, B. Soret, K. I. Pedersen, and p. Mogensen, "Ultra-reliable communications in failure-prone realistic networks," in *Proc. IEEE ISWCS*, Poznan, Poland, Sep. 2016, pp. 414–418.
[18] J. J. Nielsen, R. Liu, and P. Popovski, "Ultra-reliable low latency communication using interface diversity," *IEEE Trans. Commun.*, vol. 66, no. 3, pp. 1322–1334, Mar. 2018.
[19] R. Kotaba, R. Kotaba, C. N. Manchón, T. Balercia, and P. Popovski, "Uplink transmissions in URLLC systems with shared diversity resources," *IEEE Commun. Lett.*, to be published.
[20] J. Rao and S. Vrzic, "Packet duplication for URLLC in 5G: Architectural enhancements and performance analysis," *IEEE Netw.*, vol. 32, no. 2, pp. 32–40, Mar./Apr. 2018.
[21] 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*, Paris, France, May 2017, pp. 1005–1010.
[22] K. I. Pedersen, G. Pocovi, J. Steiner, and S. R. Khosravirad, "Punctured scheduling for critical low latency data on a shared channel with mobile broadband," in *Proc. IEEE VTC-Fall*, Toronto, ON, Canada, Sep. 2017, pp. 1–6.
[23] K. I. Pedersen, G. Pocovi, and J. Steiner, "Preemptive scheduling of latency critical traffic and its impact on mobile broadband performance," in *Proc. VTC*, Porto, Portugal, Jun. 2018, pp. 1–6.
[24] 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.
[25] S. R. Khosravirad, L. Mudolo, and K. I. Pedersen, "Flexible multi-bit feedback design for HARQ operation of large-size data packets in 5G," in *Proc. VTC Spring*, Sydney, NSW, Australia, Jun. 2017, pp. 1–5.
[26] A. A. Esswie and K. I. Pedersen, "Multi-user preemptive scheduling for critical low latency communications in 5G networks," in *Proc. IEEE ISCC*, Natal, Brazil, Jun. 2018, pp. 1–6.
[27] S. Sesia, I. Toufik, and M. Baker, "Orthogonal frequency division multiple access (OFDMA)," in *LTE—The UMTS Long Term Evolution: From Theory to Practice*. Hoboken, NJ, USA: Wiley, 2011, pp. 123–143.
[28] K. Xu, D. Tipper, Y. Qian, P. Krishnamurthy, and S. Tipmongkonsilp, "Time-varying performance analysis of multihop wireless networks with CBR traffic," *IEEE Trans. Veh. Technol.*, vol. 63, no. 7, pp. 3397–3409, Sep. 2014.
[29] J. P. Singh, Y. Li, N. Bambos, A. Bahai, B. Xu, and G. Zimmermann, "TCP performance dynamics and link-layer adaptation based optimization methods for wireless networks," *IEEE Trans. Wireless Commun.*, vol. 6, no. 5, pp. 1864–1879, May 2007.
[30] E. W. Jang, J. Lee, H. L. Lou, and J. M. Cioffi, "On the combining schemes for MIMO systems with hybrid ARQ," *IEEE Trans. Wireless Commun.*, vol. 8, no. 2, pp. 836–842, Feb. 2009.
[31] *Study on 3D Channel Model for LTE; Release 12 V12.7.0*, document TR 36.873, 3GPP, Dec. 2014.

[32] Y. Ohwatari, N. Miki, Y. Sagae, and Y. Okumura, "Investigation on interference rejection combining receiver for space–frequency block code transmit diversity in LTE-advanced downlink," *IEEE Trans. Veh. Technol.*, vol. 63, no. 1, pp. 191–203, Jan. 2014.
[33] 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.
[34] *Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures (Release 15) V15.1.0*, document TS 36.213, 3GPP, Mar. 2018.
[35] D. Bertsekas, and R. Gallager, *Data Networks*, 2nd ed. New York, NY, USA: Prentice-Hall, 1992.
[36] D. Parruca and J. Gross, "Throughput analysis of proportional fair scheduling for sparse and ultra-dense interference-limited OFDMA/LTE networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 10, pp. 6857–6870, Oct. 2016.
[37] Y. Han, S. Jin, J. Zhang, J. Zhang, and K.-K. Wong, "DFT-based hybrid beamforming multiuser systems: Rate analysis and beam selection," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 3, pp. 514–528, Jun. 2018.
[38] T. L. Narasimhan and A. Chockalingam, "Channel hardening-exploiting message passing (CHEMP) receiver in large-scale MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 847–860, Oct. 2014.

ALI A. ESSWIE received the M.Sc. degree in electrical and computer engineering from Memorial University, Canada, in 2017. He is currently pursuing the Ph.D. degree with the Department of Electronic Systems, Aalborg University. From 2013 to 2016, he was a Wireless Research Engineer with Intel Labs and Huawei Technologies, respectively. He is also with Nokia Bell Labs, Aalborg. His research interests include the 5G new radio, MAC scheduling, ultra-reliable and low latency communications, massive MIMO, and channel estimation.

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 a Research Team at Nokia Bell Labs, Aalborg. He is also 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. He is the inventor on several patents. 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 special emphasis on mechanisms that offer improved end-to-end performance delivery. He is currently a part of the EU funded Research Project ONE5G that focus on E2E-aware optimizations and advancements for the Network Edge of 5G New Radio.