Capacity Optimization of Spatial Preemptive Scheduling for Joint URLLC-eMBB Traffic in 5G New Radio

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

Published in:
2018 IEEE Globecom Workshops (GC Wkshps)

DOI (link to publication from Publisher):
10.1109/GLOCOMW.2018.8644070

Publication date:
2019

Document Version
Accepted author manuscript, peer reviewed version

Capacity Optimization of Spatial Preemptive Scheduling for Joint URLLC-eMBB Traffic in 5G New Radio

Ali A. Esswie[1,2], *Member, IEEE*, and Klaus I. Pedersen[1,2], *Senior Member, IEEE*
[1]Nokia Bell-Labs, Aalborg, Denmark
[2]Department of Electronic Systems, Aalborg University, Denmark

***Abstract*—Ultra-reliable and low-latency communication (URLLC) is envisioned as a primary service class of the fifth generation mobile networks. URLLC applications demand stringent radio latency requirements of 1 millisecond with 99.999% confidence. Obviously, the coexistence of the URLLC services and enhanced mobile broadband (eMBB) applications on the same spectrum imposes a challenging scheduling problem. In this paper, we propose an enhanced spatial preemptive scheduling framework for URLLC-eMBB traffic coexistence. The proposed scheduler ensures an instant and interference-free signal subspace for critical URLLC transmissions, while achieving best-effort eMBB performance. Furthermore, the impacted eMBB capacity is then recovered by limited network-assisted signaling. The performance of the proposed scheduler is evaluateeed by highly detailed system level simulations of the major performance indicators. Compared to the state-of-the-art multi-traffic schedulers from industry and academia, the proposed scheduler meets the stringent URLLC latency requirements, while significantly improving the achievable ergodic capacity.**

***Index Terms*— URLLC; eMBB; 5G; Preemptive scheduling, MU-MIMO, Latency.**

I. Introduction

The fifth generation (5G) of the mobile communications features two major service classes: ultra-reliable and low-latency communications (URLLC) and enhanced mobile broadband (eMBB) [1, 2]. eMBB applications support stable and delay-tolerant connections with extremely high data rates. However, URLLC critical services demand very low radio latency of 1 millisecond with 10^{-5} outage probability [3]. This category of the URLLC quality of service (QoS) is vastly different from that of the current 4G technology, where the spectral efficiency (SE) is the prime objective. Hence, the support of URLLC is envisioned to enable many future real-time applications such as virtual reality, self-driving vehicles, and tactile internet [4].

However, in pursuit of such extreme SE requirements for eMBB services and tight latency & reliability targets for URLLC, a prime scheduling challenge is how to strategically multiplex such diverse requirements on same spectrum [5]. For instance, to satisfy the URLLC latency and reliability budgets, the system must be forcibly engineered such that blocking a URLLC packet at an arbitrary transmission time interval (TTI) is a rare event. Such scheduling behavior imposes a severe degradation of the overall ergodic capacity, due to the fundamental trade-off between reliability, latency and SE [6].

In the recent open literature, eMBB and URLLC service coexistence in 5G new radio (NR) has gained progressive research attention from industry and academia. Such multi-service scheduling problem is the dominant study item of the upcoming 3GPP release-16 [7]. Furthermore, user-centric TTI scheduling is demonstrated as essential to achieve the URLLC latency and reliability targets [8, 9], i.e., URLLC users are scheduled on a short TTI duration; however, eMBB users on a longer TTI duration. Spatial diversity techniques are also considered key enablers for URLLC, to enhance the URLLC decoding ability by preserving a sufficient signal-to-interference-noise-ratio (SINR) level [10, 11]. Moreover, URLLC preemptive scheduling (PS) [12] is a state-of-the-art technique to instantly schedule sporadic URLLC traffic with minimum queuing delay. If the radio resources are monopolized by ongoing eMBB transmissions, PS scheduler immediately overwrites part of eMBB physical resource blocks (PRBs) for the sake of the incoming URLLC traffic.

In [5], we demonstrated that a standard multi-user multiple-input multiple-output (MU-MIMO) URLLC-eMBB transmission on top of PS scheduler (MUPS) is an attractive solution to provide a fair trade-off between URLLC performance and overall SE. That is, the MUPS scheduler first attempts a URLLC-eMBB MU-MIMO transmission. If the MU pairing is not possible at an arbitrary TTI, MUPS scheduler rolls back to PS scheduler for instant URLLC scheduling. However, when the system spatial degrees of freedom (SDoFs) are limited, MUPS scheduler offers a limited MU gain and degraded URLLC latency and reliability, since the standard MU-MIMO pairing condition is only constrained by the achievable sum rate. In our recent studies [13, 14], we have introduced a null space based preemptive scheduler (NSBPS), altering the MU pairing condition to instantly offer an interference-free signal subspace for sporadic URLLC traffic, through subspace projection, where the loss in the ergodic capacity is upper-bounded by the eMBB projection loss.

In this paper, an enhanced NSBPS (eNSBPS) scheduling framework for downlink (DL) 5G-NR is proposed. When incoming URLLC traffic can not be immediately scheduled, i.e., without queuing or segmentation, the eNSBPS scheduler immediately alters the system optimization to a region where the URLLC QoS is instantly guaranteed, and delay-tolerant eMBB QoS is recovered through limited network-assisted signaling. eNSBPS searches for an active eMBB transmission whose transmission is most aligned within a pre-defined reference spatial subspace. Next, eNSBPS projects the selected

eMBB transmission onto the reference subspace for which its instantly paired URLLC user, on the same resources, aligns its decoding matrix into a possible null space of the reference subspace; thus, experiencing an interference-free transmission. Then. the base-station (BS) signals the victim eMBB users with limited signaling components in the control channel to recover the inflicted capacity loss due to the instant spatial projection, hence, achieving the maximum possible ergodic capacity of a multi-traffic MU system. Compared to the state-of-the-art scheduler proposals, eNSBPS scheduling framework shows a robust URLLC performance with a significantly improved ergodic capacity.

Due to the complexity of the 5G-NR system model and addressed problems herein [1-3], the performance of the proposed eNSBPS scheduler is validated using highly detailed system level simulations (SLSs), where the majority of the 5G-NR protocol stack is implemented and calibrated against the 3GPP 5G-NR assumptions, including but not limited to: dynamic link adaptation & user scheduling, hybrid automatic repeat request (HARQ), 3D channel modeling and estimation.

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 $\|\mathcal{X}\|$ are the mean and 2-norm of $\mathcal{X}$. $\angle\mathcal{X}$ denotes the principal phase direction of $\mathcal{X}$. $\mathcal{X}^{\kappa}, \kappa\in\{\mathrm{llc}, \mathrm{mbb}\}$ denotes the type of user $\mathcal{X}$, $\mathbb{E}\{\mathcal{X}\}$ and $\mathbf{card}(\mathcal{X})$ are the statistical expectation and cardinality of $\mathcal{X}$.

This paper is organized as follows. Section II presents the system model of this work. Section III introduces the problem formulation and detailed description of the proposed scheduler framework. Section IV discusses the performance evaluation results. The paper is concluded in Section V while work acknowledgments are presented in Section VI.

II. System Model

We adopt a DL MU-MIMO transmission in 5G-NR [13, 14], where there are C cells, each with N_t transmit antennas. Each cell serves $K_{\mathrm{mbb}}+K_{\mathrm{llc}}=K$ users on average, each with M_r receive antennas, where K_{mbb} and K_{llc} are the average numbers of eMBB and URLLC users per cell. We asses two types of DL traffic as: a) URLLC sporadic FTP3 traffic with B−byte payload size and a Poisson point arrival λ, and b) eMBB full buffer traffic with infinite payload size. As depicted in Fig. 1, the agile 5G-NR frame structure is considered in this work, where URLLC traffic is scheduled on a short TTI duration to satisfy its stringent latency targets, i.e., 2-symbol TTI, while eMBB users can be scheduled on a longer TTI duration, i.e., 14-symbol TTI, to maximize the achievable SE. In the frequency domain, users are dynamically multiplexed using orthogonal frequency division multiple access, where the smallest schedulable unit is the PRB, i.e., 12 sub-carriers of 15 kHz sub-carrier spacing.

We assume a maximum subset $\mathtt{G}_c\in\mathcal{K}_c$ of MU URLLC-eMBB co-scheduled users over an arbitrary PRB in the c^{th} cell, where $G_c=\mathbf{card}(\mathtt{G}_c)$, $G_c\leq N_t$ is the MU user rank per PRB and $\mathcal{K}_c$ is the set of active eMBB/URLLC users in the c^{th} cell. The DL received signal at the k^{th} user from the c^{th} cell is expressed by

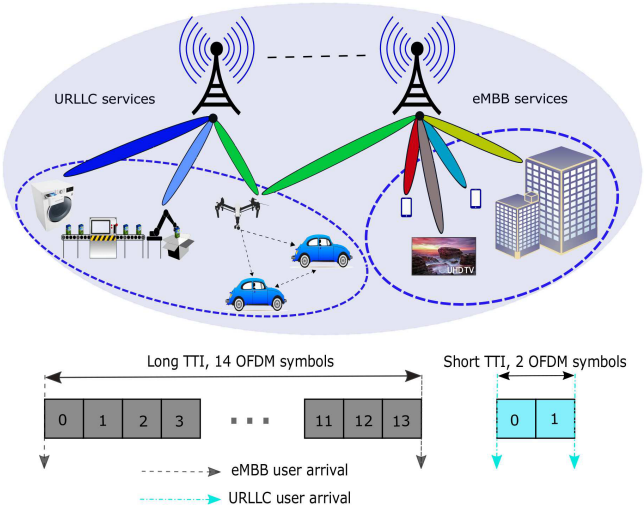

Fig. 1. Agile TTI structure in 5G-NR.

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

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 spatial channel matrix [15]. $\mathbf{v}^{\kappa}_{k,c}\in\mathcal{C}^{N_t\times 1}$ is the zero-forcing beamforming vector, with a single spatial stream per user, and is calculated as: $\mathbf{v}^{\kappa}_{k,c}=\left(\mathbf{H}^{\kappa}_{k,c}\right)^{\mathrm{H}}\left(\mathbf{H}^{\kappa}_{k,c}\left(\mathbf{H}^{\kappa}_{k,c}\right)^{\mathrm{H}}\right)^{-1}$. Finally, $s^{\kappa}_{k,c}$ and $\mathbf{n}^{\kappa}_{k,c}$ indicate the transmitted symbol and the additive white Gaussian noise, respectively. The first summation represents the intra-cell inter-user interference while the latter introduces the inter-cell interference, resulting from the URLLC and eMBB traffic. The received signal is then decoded by the linear minimum mean square interference rejection combining (LMMSE-IRC) [16] vector $\mathbf{u}^{\kappa}_{k,c}$ as

$$\left(\mathrm{y}^{\kappa}_{k,c}\right)^{*}=\left(\mathbf{u}^{\kappa}_{k,c}\right)^{\mathrm{H}}\mathrm{y}^{\kappa}_{k,c}, \quad (2)$$

where $\left(\mathrm{y}^{\kappa}_{k,c}\right)^{*}$ is the post-combining received signal. Then, the received SINR at the k^{th} user can be represented by:

$$\Upsilon^{\kappa}_{k,c}=\frac{p^{c}_{k}\left\|\mathbf{H}^{\kappa}_{k,c}\mathbf{v}^{\kappa}_{k,c}\right\|^{2}}{1+\sum\limits_{g\in\mathtt{G}_c,g\neq k}p^{c}_{g}\left\|\mathbf{H}^{\kappa}_{k,c}\mathbf{v}^{\kappa}_{g,c}\right\|^{2}+\sum\limits_{j\in C,j\neq c}\sum\limits_{g\in\mathtt{G}_j}p^{j}_{g}\left\|\mathbf{H}^{\kappa}_{g,j}\mathbf{v}^{\kappa}_{g,j}\right\|^{2}}, \quad (3)$$

where p^{c}_{k} represents the transmit power towards the k^{th} user.

III. Proposed eNSBPS Scheduler

A. Problem Formulation

Inline with the 5G-NR targets in the upcoming 3GPP release-16 [7], the eMBB and URLLC QoS classes have to be efficiently multiplexed on the same spectrum. Such requirement implies that the QoS objective functions of the MAC scheduler should be user-centric, instead of network-centric [14]. However, these QoS classes are highly correlated

and need to be reliably satisfied, e.g., eMBB SE maximization and URLLC latency minimization as

$$\forall k_{\text{mbb}} \in \mathcal{K}_{\text{mbb}} : R_{\text{mbb}} = \underset{k_{\text{mbb}} \in \mathcal{K}_{\text{mbb}}}{\arg\max} \sum_{k_{\text{mbb}}=1}^{K_{\text{mbb}}} \sum_{rb \in \Xi^{\text{mbb}}_{k_{\text{mbb}}}} \beta_{k_{\text{mbb}}} r^{\text{mbb}}_{k_{\text{mbb}},rb}, \quad (4)$$

$$\forall k_{\text{llc}} \in \mathcal{K}_{\text{llc}} : \underset{k_{\text{llc}} \in \mathcal{K}_{\text{llc}}}{\arg\min} \left(\Psi_{k_{\text{llc}}} \right), \Psi_{k_{\text{llc}}} \leq 1\,\text{ms}, \quad (5)$$

where R_{mbb} is the overall eMBB ergodic capacity, $\mathcal{K}_{\text{mbb}}$ and $\mathcal{K}_{\text{llc}}$ are the active sets of eMBB and URLLC users, respectively. $\Xi^{\text{mbb}}_{k_{\text{mbb}}}$ and $\gamma_{k_{\text{mbb}}}$ are the set of granted PRBs and the scheduling priority of the k^{th}_{mbb} user, respectively, $r^{\text{mbb}}_{k_{\text{mbb}},rb}$ is the achievable per-PRB rate of the k^{th}_{mbb} user. Finally, $\Psi_{k_{\text{llc}}}$ denotes the URLLC one-way radio latency, which can be expressed as (assuming a successful first transmission):

$$\Psi_{k_{\text{llc}}} = \Lambda_{\text{q}} + \Lambda_{\text{bsp}} + \Lambda_{\text{fa}} + \Lambda_{\text{tx}} + \Lambda_{\text{uep}}, \quad (6)$$

where $\Lambda_{\text{q}}, \Lambda_{\text{bsp}}, \Lambda_{\text{fa}}, \Lambda_{\text{tx}}, \Lambda_{\text{uep}}$ are random variables to present the URLLC queuing, BS processing, frame alignment, transmission, and user processing delays, respectively. Due to the agile 5G-NR frame structure, Λ_{fa} is upper-bounded by a short TTI duration while Λ_{bsp} & Λ_{uep} are each bounded by 3-OFDM symbol duration [17], due to the enhanced processing capabilities that come with the 5G-NR. Thus, the URLLC queuing delay Λ_{q} and transmission delay Λ_{tx} are the major bottleneck against achieving the stringent URLLC latency targets. As reported in our recent studies [13, 14], these delay components are hardly controlled in a dynamic system, and highly correlated to each others. Furthermore, their statistical behavior vastly varies with the URLLC arrival rate λ, packet size B, SINR level $\Upsilon^{\text{llc}}_{k_{\text{llc}}}$, and the scheduler buffering behavior.

To achieve the URLLC stringent latency and reliability requirements in eq. (5), Λ_{q} and Λ_{tx} must be always controlled at minimum to allow for further delay allowance for the re-transmission(s) within the target 1 ms. This can only be achieved by enforcing a hard URLLC priority in the scheduler queues, or allocating URLLC users with excessive PRB sizes to ensure a sufficient outage SINR level. In both cases, the eMBB utility function in eq. (4) is severely under-optimized, resulting in a significant degradation of the system ergodic capacity. In this work, we address such challenging multiplexing requirement and propose a scheduling framework that guarantees the URLLC QoS while significantly improving the system SE.

B. Proposed eNSBPS Scheduler – At The BS Side

During an arbitrary TTI, eNSBPS scheduler assigns single-user (SU) resources to new/buffered eMBB traffic, if there are no new URLLC arrivals, based on the proportional fair (PF) [18] criterion as

$$\Theta\left\{\text{PF}_{k_{\text{mbb}}}\right\} = \frac{r^{\text{mbb}}_{k_{\text{mbb}},rb}}{\overline{r}^{\text{mbb}}_{k_{\text{mbb}},rb}}, \quad (7)$$

$$k^{*}_{\text{mbb}} = \underset{k_{\text{mbb}} \in \mathcal{K}_{\text{mbb}}}{\arg\max} \ \Theta\left\{\text{PF}_{k_{\text{mbb}}}\right\}, \quad (8)$$

where $\overline{r}^{\text{mbb}}_{k_{\text{mbb}},rb}$ is the mean delivered data rate of the k^{th}_{mbb} user. Though, if there are new/buffered URLLC packets at the BS queues, and the instant schedulable resources are sufficiently enough to accommodate such payloads, the eNSBPS scheduler overwrites the SU eMBB scheduling priority for the sake of the URLLC traffic, by applying the weighted PF (WPF) criterion as

$$\Theta\left\{\text{WPF}_{k_{\kappa}}\right\} = \frac{r^{\kappa}_{k_{\kappa},rb}}{\overline{r}^{\kappa}_{k_{\kappa},rb}} \gamma_{k_{\kappa}}, \quad (9)$$

with $\gamma_{k_{\text{llc}}} \gg \gamma_{k_{\text{mbb}}}$ for immediate URLLC SU scheduling. In case radio resource are not immediately sufficient for the incoming URLLC packets, the eNSBPS scheduler first attempts a highly conservative version of a standard MU-MIMO transmission between the URLLC-eMBB user pair. That is, users are only paired if their corresponding transmission subspaces offer high spatial separation [14] as

$$1 - \left| \left(\mathbf{v}^{\text{mbb}}_{k_{\text{mbb}}} \right)^{\text{H}} \mathbf{v}^{\text{llc}}_{k_{\text{llc}}} \right|^2 \geq \eta, \quad (10)$$

where $\eta \rightarrow [0, 1]$ is a conservative orthogonality threshold. However, if such orthogonality can not be offered at an arbitrary TTI, due to limited SDoFs, the proposed eNSBPS instantly enforces such orthogonality, for the sake of the URLLC traffic. It pre-defines a discrete Fourier transform spatial reference subspace, pointing towards an arbitrary direction θ as given by

$$\mathbf{v}_{\text{ref}}(\theta) = \left(\frac{1}{\sqrt{N_t}} \right) \left[1, e^{-j2\pi\Delta\cos\theta}, \ldots, e^{-j2\pi\Delta(N_t-1)\cos\theta} \right]^{\text{T}}, \quad (11)$$

where Δ is the antenna spacing. Then, scheduler instantly searches for an active eMBB whose transmission is most aligned within the reference subspace, using the minimum Chordal distance as

$$k^{\diamond}_{\text{mbb}} = \arg \underset{\mathcal{K}_{\text{mbb}}}{\min} \mathbf{d} \left(\mathbf{v}^{\text{mbb}}_{k}, \mathbf{v}_{\text{ref}} \right), \quad (12)$$

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

$$\mathbf{d} \left(\mathbf{v}^{\text{mbb}}_{k}, \mathbf{v}_{\text{ref}} \right) = \frac{1}{\sqrt{2}} \left\| \mathbf{v}^{\text{mbb}}_{k} \left(\mathbf{v}^{\text{mbb}}_{k} \right)^{\text{H}} - \mathbf{v}_{\text{ref}} \mathbf{v}^{\text{H}}_{\text{ref}} \right\|. \quad (13)$$

Finally, the eNSBPS scheduler instantly projects the selected eMBB transmission onto the reference subspace as:

$$\left(\mathbf{v}^{\text{mbb}}_{k^{\diamond}_{\text{mbb}}} \right)' = \frac{\mathbf{v}^{\text{mbb}}_{k^{\diamond}_{\text{mbb}}} \cdot \mathbf{v}_{\text{ref}}}{\left\| \mathbf{v}_{\text{ref}} \right\|^2} \times \mathbf{v}_{\text{ref}}, \quad (14)$$

with $\left(\mathbf{v}^{\text{mbb}}_{k^{\diamond}_{\text{mbb}}} \right)'$ as the post-projection eMBB precoder. As shown in Fig. 2, the victim eMBB transmission inflicts a loss in its principal direction and gain, respectively, due to the instant projection at the BS, as it will be discussed in greater detail in Section III-D. Then, scheduler immediately allocates shared resources between the incoming URLLC user and the victim eMBB transmission. Finally, as depicted by the timing diagram in Fig. 3, the BS signals the URLLC user by a single-bit true co-scheduling indication, i.e., $\alpha = 1$ in the control channel, for the URLLC user to de-orient its decoding matrix into

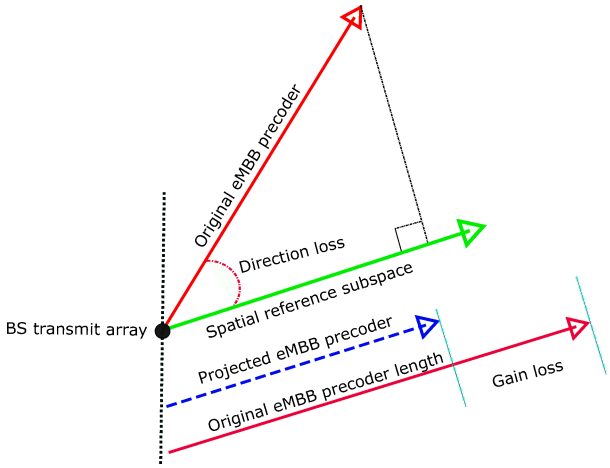

Fig. 2. Victim eMBB transmission projection.

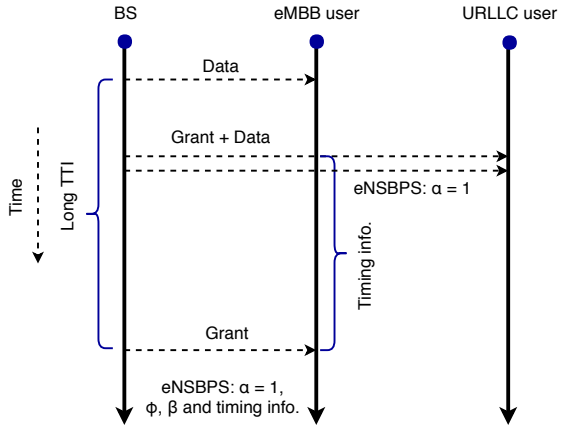

Fig. 3. Timing diagram of the eNSBPS scheduler.

one possible null space of the reference subspace, hence, experiencing no intra-cell inter-user interference. Furthermore, to recover the eMBB capacity region, being impacted by the instant spatial projection, the BS also signals the victim eMBB user with:

- $\alpha = 1$, AND
- Multi-bit separation angle $\Phi = \left| \angle \left(\mathbf{v}_{k^{\diamond}_{\text{mbb}}}^{\text{mbb}} \right) - \angle \left(\mathbf{v}_{\text{ref}} \right) \right|$ between its original principal precoder and the reference subspace, AND
- Timing information of the starting symbol when such spatial projection has been applied, AND/OR
- Multi-bit original precoder length $\beta = \left\| \mathbf{v}_{k^{\diamond}_{\text{mbb}}}^{\text{mbb}} \right\|$.

C. Proposed eNSBPS Scheduler – At The URLLC User Side

Upon the reception of a true co-scheduling indication $\alpha = 1$ in the control channel, the URLLC user realizes that its scheduling grant is shared with an active eMBB user, whose transmission is aligned within the pre-known reference subspace. The URLLC user first designs its first-stage LMMSE-IRC decoding matrix in order to reject the inter-cell interference statistics as

$$\left(\mathbf{u}_k^{\text{llc}} \right)^{(1)} = \left(\mathbf{H}_k^{\text{llc}} \mathbf{v}_k^{\text{llc}} \left(\mathbf{H}_k^{\text{llc}} \mathbf{v}_k^{\text{llc}} \right)^{\text{H}} + \mathbf{W} \right)^{-1} \mathbf{H}_k^{\text{llc}} \mathbf{v}_k^{\text{llc}}, \tag{15}$$

with the interference covariance matrix $\mathbf{W}$ given as

$$\mathbf{W} = \mathbb{E} \left(\mathbf{H}_k^{\text{llc}} \mathbf{v}_k^{\text{llc}} \left(\mathbf{H}_k^{\text{llc}} \mathbf{v}_k^{\text{llc}} \right)^{\text{H}} \right) + \sigma^2 \mathrm{I}_{M_r}, \tag{16}$$

where I_{M_r} is $M_r \times M_r$ identity matrix. Then, $\left(\mathbf{u}_k^{\text{llc}} \right)^{(1)}$ is transferred into one possible null space of the inter-user interference effective channel $\mathbf{H}_k^{\text{llc}} \mathbf{v}_{\text{ref}}$, coming from the paired eMBB user and aligned within the reference subspace as

$$\left(\mathbf{u}_k^{\text{llc}} \right)^{(2)} = \left(\mathbf{u}_k^{\text{llc}} \right)^{(1)} - \frac{\left(\left(\mathbf{u}_k^{\text{llc}} \right)^{(1)} \cdot \mathbf{H}_k^{\text{llc}} \mathbf{v}_{\text{ref}} \right)}{\left\| \mathbf{H}_k^{\text{llc}} \mathbf{v}_{\text{ref}} \right\|^2} \times \mathbf{H}_k^{\text{llc}} \mathbf{v}_{\text{ref}}. \tag{17}$$

This way, the second-stage decoder $\left(\mathbf{u}_k^{\text{llc}} \right)^{(2)}$ matrix of the URLLC user experiences no inter-user interference, boosting its received SINR level.

D. Proposed eNSBPS Scheduler – At The eMBB User Side

At the eMBB user side, when $\alpha = 1$ is received, it acknowledges that its corresponding transmission is being spatially altered *on-the-fly* to be aligned within the reference subspace. Thus, it inflicts a spatial loss in its spatial gain and principal direction, respectively, e.g., as described in Fig. 2 and eq. (14), the loss in the precoding spatial gain is given by

$$\left\| \left(\mathbf{v}_{k^{\diamond}_{\text{mbb}}}^{\text{mbb}} \right)' \right\| = \left\| \mathbf{v}_{\text{ref}} \right\| \left\| \mathbf{v}_{k^{\diamond}_{\text{mbb}}}^{\text{mbb}} \right\| \cos\left(\Phi \right), \tag{18}$$

where $\left\| \mathbf{v}_{\text{ref}} \right\| = 1$, and the original precoder spatial length $\left\| \mathbf{v}_{k^{\diamond}_{\text{mbb}}}^{\text{mbb}} \right\|$ exhibits a scale-down loss by $\cos\left(\Phi \right)$. Thus, we introduce two setups to recover the eMBB capacity with different signaling overhead as follows.

Setup-1: victim eMBB user attempts to reconstruct its original transmission subspace, that was altered at the BS by the instant spatial projection, and based on the knowledge of the reference subspace, Φ, and β, expressed as

$$\left(\mathbf{v}_{k^{\diamond}_{\text{mbb}}}^{\text{mbb}} \right)^{\text{est.}} = \beta \, e^{-j\Phi} \, \mathbf{v}_{\text{ref}}, \tag{19}$$

where $\left(\mathbf{v}_{k^{\diamond}_{\text{mbb}}}^{\text{mbb}} \right)^{\text{est.}}$ is the estimated original transmission subspace of the victim eMBB user. The first factor β compensates for the loss in the precoder spatial length; however, the second factor $e^{-j\Phi}$ cancels the spatial rotation effect. Then, the eMBB user projects its first-stage LMMSE-IRC decoding matrix $\left(\mathbf{u}_k^{\text{mbb}} \right)^{(1)}$ on its desired estimated effective transmission subspace $\mathbf{H}_k^{\text{mbb}} \left(\mathbf{v}_{k^{\diamond}_{\text{mbb}}}^{\text{mbb}} \right)^{\text{est.}}$ as

$$\left(\mathbf{u}_k^{\text{mbb}} \right)^{(2)} = \frac{\left(\left(\mathbf{u}_k^{\text{mbb}} \right)^{(1)} \cdot \mathbf{H}_k^{\text{mbb}} \left(\mathbf{v}_{k^{\diamond}_{\text{mbb}}}^{\text{mbb}} \right)^{\text{est.}} \right)}{\left\| \mathbf{H}_k^{\text{mbb}} \left(\mathbf{v}_{k^{\diamond}_{\text{mbb}}}^{\text{mbb}} \right)^{\text{est.}} \right\|^2} \times \mathbf{H}_k^{\text{mbb}} \left(\mathbf{v}_{k^{\diamond}_{\text{mbb}}}^{\text{mbb}} \right)^{\text{est.}}, \tag{20}$$

with $\left(\mathbf{u}_k^{\text{mbb}} \right)^{(2)}$ as the second-stage eMBB decoder, that is de-oriented towards its original transmission subspace, thus, maximizing its achievable capacity.

Setup-2: based on the fact that both the length and direction loss of the victim eMBB user depend on the spatial separation angle between its original precoder and the reference subspace, i.e., spatial rotation of Φ, and spatial gain loss factor of

Table I
SIMULATION PARAMETERS.

Parameter	Value
Environment	3GPP-UMA,7 gNBs, 21 cells, 500 meters inter-site distance
Channel bandwidth	10 MHz, FDD
Antenna setup	BS: 8 Tx, UE: 2 Rx
User dropping	uniformly distributed URLLC: 7 users/cell eMBB: 7 users/cell
User receiver	LMMSE-IRC
TTI configuration	URLLC: 0.143 ms (2 OFDM symbols) eMBB: 1 ms (14 OFDM symbols)
CQI	periodicity: 5 ms, with 2 ms latency
HARQ	asynchronous HARQ, Chase combining HARQ round trip time = 4 TTIs
Link adaptation	dynamic modulation and coding target URLLC BLER : 1% target eMBB BLER : 10%
Traffic model	URLLC: bursty, B=50 bytes, $\lambda = 250$ eMBB: full buffer

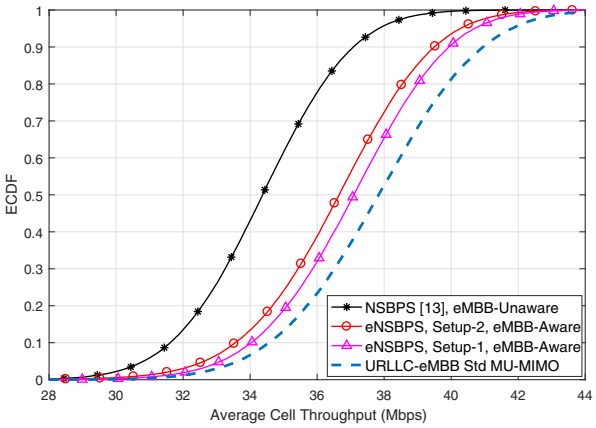

Fig. 4. Average cell throughput performance (Mbps).

$\cos(\Phi)$. Thus, the signaling overhead from the BS to the intended eMBB users can only be limited to Φ, without the need for signaling β. Accordingly, a spatial rotation matrix Γ is constructed and scaled-up by the length loss factor as

$$\Gamma = \left(\frac{1}{\cos(\Phi)}\right) \begin{bmatrix} \left(e^{-j\Phi}\right)_{0,0} & \cdots & \left(e^{-j\Phi}\right)_{0,d-1} \\ \vdots & \ddots & \vdots \\ \left(e^{-j\Phi}\right)_{M_r-1,0} & \cdots & \left(e^{-j\Phi}\right)_{M_r-1,d-1} \end{bmatrix}, \tag{21}$$

where d indicates the number of spatial streams per user. Finally, inline with setup-1, the victim eMBB user projects its fist-stage decoding matrix onto the spatial rotation matrix, given by

$$\left(\mathbf{u}_k^{\text{mbb}}\right)^{(2)} = \frac{\left(\left(\mathbf{u}_k^{\text{mbb}}\right)^{(1)} \cdot \Gamma\right)}{\|\Gamma\|^2} \times \Gamma. \tag{22}$$

IV. SIMULATION RESULTS

In this section, we introduce the SLS results of the proposed eNSBPS scheduler, following the 5G-NR assumptions [5]. The major simulation parameters are listed in Table I.

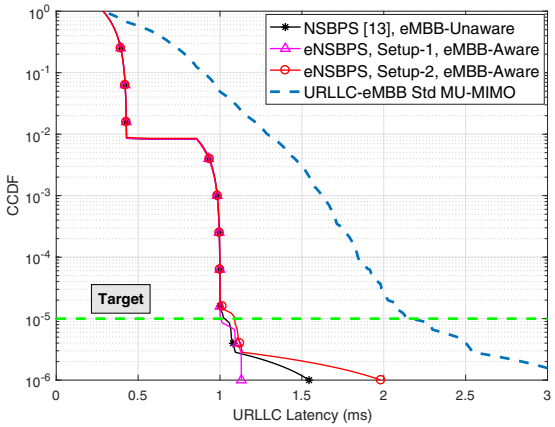

Fig. 5. URLLC one-way latency performance (ms).

We present a performance comparison of three state-of-the-art schedulers for joint eMBB and URLLC traffic as follows: (1) proposed eNSBPS scheduler with the two techniques to recover the impacted eMBB capacity, (2) our recent NSBPS scheduler [13], where the victim eMBB users are presumed unaware of the spatial projection, hence, a degraded eMBB capacity is exhibited, and (3) a standard (Std) MU-MIMO scheduler between incoming URLLC users and ongoing eMBB transmissions, if the instantly available resources are not sufficient to accommodate the entire URLLC payload.

Fig. 4 presents the empirical cumulative distribution function (ECDF) of the average achievable cell throughput in Mbps of all three schedulers under evaluation. As can be noticed, the Std URLLC-eMBB MU-MIMO scheduler offers the maximum possible cell throughput since the eMBB transmissions are not biasedly altered for the sake of the URLLC traffic; however, with the worst URLLC latency performance as will be shown in Fig. 5. The proposed eNSBPS scheduler with the two introduced eMBB recovery techniques, significantly improves the achievable cell throughput against the eMBB-unaware NSBPS scheduler. It approaches the Std MU-MIMO scheduler, while simultaneously preserving the URLLC latency targets. This is because the intentionally lost eMBB capacity at the BS is recovered at the victim eMBB users using BS control signaling. Both setup-1 and setup-2 of the proposed eNSBPS scheduler show a similar cell throughput performance, with further reduced signaling overhead for setup-2, since both the spatial length and direction losses of the victim eMBB users only depend on the separation angle between the eMBB original precoder and the reference subspace at the BS.

Examining the URLLC performance, Fig. 5 depicts the complementary CDF (CCDF) of the URLLC one-way latency in ms. As can be clearly identified, both proposed eNSBPS and NSBPS schedulers achieve the stringent URLLC latency target of 1 ms at 10^{-5} outage, since under both schedulers, sporadic URLLC traffic is guaranteed an instant and interference-free spatial subspace, hence, improving the URLLC decoding ability and reducing the number of inflicted URLLC re-transmissions. Furthermore, due to the fact that the Std MU-MIMO pairing condition is only constrained by the achievable sum rate, i.e., not a user-centric constraint,

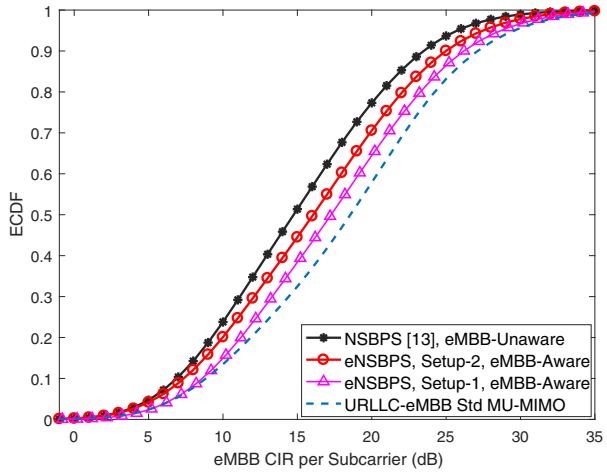

Fig. 6. Average eMBB CIR per sub-carrier performance (dB).

a Std URLLC-eMBB MU-MIMO transmission degrades the URLLC decoding SINR level. Additionally, a Std MU-MIMO pairing is not almost surely guaranteed, e.g., if the SDoFs are limited during an arbitrary TTI, MU pairing may not be possible, hence, the incoming URLLC traffic must be queued for multiple TTIs until sufficient radio resources are released. As a result, the Std URLLC-eMBB MU-MIMO scheduler exhibits a significant loss of the URLLC latency performance, not fulfilling its latency targets.

Finally, looking at the individual eMBB performance, Fig. 6 presents the ECDF of the eMBB user post-detection carrier-to-interference-ratio (CIR) in dB. The Std MU-MIMO scheduler offers the best eMBB CIR since the paired eMBB users are only impacted by the standard MU equal-power sharing and the resultant inter-user interference. That is, eMBB transmissions are not spatially altered for the sake of the paired URLLC traffic, leading to a better cell performance as shown in Fig. 4. On another side, NSBPS scheduler suffers from the worst eMBB CIR due to the unrealizable eMBB projections. Hence, victim eMBB users exhibit a sub-optimal LMMSE-IRC performance since both the actual and estimated eMBB effective channels are not aligned within the same signal subspace. The proposed eNSBPS, under the two introduced recovery setups, provides a clear enhancement of the end eMBB CIR performance. The eMBB recovery mechanisms of the eNSBPS scheduler re-align the LMMSE-IRC decoding spatial span of the victim eMBB users into its original signal subspace before the inflicted projection at the BS, thus, maximizing their perceived effective channels and SNR levels, respectively.

V. Conclusion

In this work, an enhanced null space based preemptive scheduler (eNSBPS) has been introduced for joint URLLC and eMBB traffic in 5G new radio. Sporadic URLLC traffic is instantly guaranteed an interference-free subspace for immediate and secured transmission without queuing, through eMBB subspace projection. Thus, proposed eNSBPS scheduler offers extreme URLLC latency robustness. The impacted eMBB capacity is then recovered through subspace alignment at the victim eMBB users, hence, maximizing the achievable eMBB capacity. Compared to the state-of-the-art scheduling proposals, extensive system level simulations show that proposed scheduler framework satisfies the stringent URLLC latency targets while significantly improving the overall cell spectral efficiency, by achieving an average gain of ~ 3.2 dB in the eMBB post-detection carrier-to-interference-ratio.

VI. Acknowledgments

This work is partly funded by the Innovation Fund Denmark (IFD) – case number: 7038-00009B. Also, part of this work has been performed in the framework of the Horizon 2020 project ONE5G (ICT-760809) receiving funds from the European Union. 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] Study on new radio access technology (Release 14), 3GPP, TR 38.801, V14.0.0, March 2017.

[2] NR and NG-RAN overall description; Stage-2 (Release 15), 3GPP, TS 38.300, V2.0.0, Dec. 2017.

[3] IMT vision – "Framework and overall objectives of the future development of IMT for 2020 and beyond", international telecommunication union (ITU), ITU-R M.2083-0, Feb. 2015.

[4] M. Simsek, A. Aijaz, M. Dohler, J. Sachs, and G. P. Fettweis, "5G-enabled tactile internet", *in IEEE J. Sel. Areas Commun.*, vol. 34, no. 3, pp. 460- 473, Mar. 2016.

[5] Ali A. Esswie, and K.I. Pedersen, "Multi-user preemptive scheduling for critical low latency communications in 5G networks," *in Proc. IEEE ISCC*, Natal, 2018, pp. 1-6.

[6] 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, 2014, pp. 1391-1396.

[7] Study on enhancement of URLLC supporting in 5GC (Work item: release-16); 3GPP, TR *to be specified*, March 2018.

[8] 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.

[9] G. Pocovi, H. Shariatmadari, G. Berardinelli, K. Pedersen, J. Steiner and Z. Li, "Achieving URLLC: challenges and envisioned system enhancements," *IEEE Netw.*, vol. 32, no. 2, pp. 8-15, April 2018.

[10] G. Pocovi, B. Soret, M. Lauridsen, K. I. Pedersen and P. Mogensen, "Signal quality outage analysis for URLLC in cellular networks," *in Proc. IEEE Globecom*, San Diego, CA, 2015, pp. 1-6.

[11] J. J. Nielsen, R. Liu and P. Popovski, "URLLC using interface diversity," *IEEE Trans. Commun.*, vol. 66, no. 3, pp. 1322-1334, March 2018.

[12] 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, 2018, pp. 1-6.

[13] Ali A. Esswie, and K.I. Pedersen, "Null space based preemptive scheduling for joint URLLC and eMBB traffic in 5G networks," *submitted to IEEE Globecom*, Abu Dhabi, 2018.

[14] Ali A. Esswie, and K.I. Pedersen, "Opportunistic spatial preemptive scheduling for URLLC and eMBB coexistence in multi-user 5G networks," *submitted to IEEE Netw.*, June, 2018.

[15] Study on 3D channel model for LTE; Release 12, 3GPP, TR 36.873, V12.7.0, Dec. 2014.

[16] 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.

[17] Physical layer procedures; Evolved universal terrestrial radio access (Release 15), 3GPP, TS 36.213, V15.1.0, March. 2018.

[18] 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.