Null Space Based Preemptive Scheduling For Joint URLLC and eMBB Traffic in 5G Networks

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.2018.8644351

Publication date:
2019

Document Version
Accepted author manuscript, peer reviewed version

Null Space Based Preemptive Scheduling For Joint URLLC and eMBB Traffic in 5G Networks

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*—In this paper, we propose a null-space-based preemptive scheduling framework for cross-objective optimization to always guarantee robust URLLC performance, while extracting the maximum possible eMBB capacity. The proposed scheduler perpetually grants incoming URLLC traffic a higher priority for instant scheduling. In case that radio resources are not immediately schedulable, proposed scheduler forcibly enforces an artificial spatial user separation, for the URLLC traffic to get instantly scheduled over shared resources with ongoing eMBB transmissions. A pre-defined reference spatial subspace is constructed for which scheduler instantly picks the active eMBB user whose precoder is the closest possible. Then, it projects the eMBB precoder *on-the-go* onto the reference subspace, in order for its paired URLLC user to orient its decoder matrix into one possible null space of the reference subspace. Hence, a robust decoding ability is always preserved at the URLLC user, while cross-maximizing the ergodic capacity. Compared to the state-of-the-art proposals from industry and academia, proposed scheduler shows extreme URLLC latency robustness with significantly improved overall spectral efficiency. Analytical analysis and extensive system level simulations are presented to support paper conclusions.**

***Index Terms*— URLLC; eMBB; Null space; MU-MIMO; 5G; Preemptive; Puncture scheduling.**

I. Introduction

Emerging fifth generation (5G) systems are envisioned to support two major service classes: ultra-reliable low-latency communications (URLLC) and enhanced mobile broadband (eMBB) [1]. URLLC refer to the future services that demand extremely reliable and low latency data communication, i.e., one-way latency up to 1 ms with 10^{-5} outage probability [2]. That is, the quality of service (QoS) of the URLLC-type applications is infringed if more than one packet out of 10^5 packets are not successfully decoded within the 1 ms deadline. This URLLC QoS is immensely different from that of the current long term evolution (LTE) technology [3], where the overall spectral efficiency (SE) is the prime objective.

To satisfy such stringent latency requirements, the system should be always engineered so that blocking a URLLC packet is a very rare event. Therefore, URLLC services must satisfy their individual outage capacity, instead of the ergodic capacity. That is, by setting an ultra-tight target block error rate (BLER) to always ensure a sufficient URLLC decoding ability. This way, it leads to a significant loss of the network SE due to the fundamental tradeoff between reliability, latency and the achievable SE [4].

In the recent literature, diverse 5G scheduling contributions have been introduced. User-centric scheduling with variable transmission time intervals (TTIs) [5] is essential to minimize the URLLC frame alignment and queuing delays. Furthermore, URLLC spatial diversity techniques are vital to preserve a sufficient URLLC signal-to-interference-noise-ratio (SINR). For example, the work in [6] demonstrates that a 4×4 multi-input multi-output (MIMO) microscopic diversity and two orders of macroscopic diversity are required to reach the URLLC outage SINR level. A recent study [7] further extends the usage of the spatial diversity for URLLC by flexibly allocating coded segments of the URLLC payload message to different interfaces. Thus, a better latency-reliability tradeoff can be achieved by reducing the original payload transmission time. Additionally, URLLC punctured scheduling (PS) [8] is a state-of-the-art scheme to further minimize the queuing delay of the URLLC traffic, where sporadic URLLC traffic is instantly scheduled by overwriting part of the radio resources, monopolized by ongoing eMBB transmissions.

However, the majority of the URLLC scheduling studies consider a monotonic optimization structure of the URLLC outage capacity. Therefore, URLLC requirements can be proportionally satisfied only with the size of the URLLC granted resources or received SINR levels. However, when joint eMBB and URLLC traffic coexists on the same radio spectrum, this approach fails to reach a proper system ergodic capacity.

In this work, a null-space-based preemptive scheduling (NSBPS) for joint eMBB and URLLC traffic is proposed. Proposed scheduler seeks to dynamically fulfill a jointly constrained objective, for which the URLLC QoS is guaranteed, while achieving the best possible eMBB capacity. If the available radio resources are not sufficient to accommodate the URLLC payload, NSBPS forcibly fits the URLLC traffic within an ongoing eMBB transmission in an instant, controlled, semi-transparent and biased multi-user MIMO (MU-MIMO) transmission. The proposed NSBPS instantly selects an active eMBB user whose transmission is most aligned within an arbitrary reference subspace. It spatially projects the selected eMBB transmission onto the reference subspace for which its paired URLLC user de-orients its decoding matrix into one possible null-space. Accordingly, a robust SINR level is preserved at the URLLC user side. Compared to the state-of-the-art studies, proposed NSBPS shows extreme robustness of the URLLC QoS with significantly improved ergodic capacity.

Due to the complexity of the 5G new radio (NR) system model [1-3] and addressed problems therein, the performance of the proposed scheduler is validated by extensive system simulations (SLS), and supported by analytical analysis of

the major performance indicators. Those simulations are based on widely accepted models and calibrated against the 5G NR specifications to ensure highly reliable statistical results.

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 $\mathcal{X}$. $\mathcal{X}\sim\mathbb{CN}(0,\sigma^2)$ indicates a complex Gaussian random variable with zero mean and variance σ^2, $\mathcal{X}_\kappa, \kappa\in\{\text{llc},\text{mbb}\}$ denotes the type of user $\mathcal{X}$, $\mathbb{E}\{\mathcal{X}\}$ and $\mathbf{card}(\mathcal{X})$ are the statistical expectation and cardinality of $\mathcal{X}$.

The paper is organized as follows. Section II presents the system and signal models, respectively. Section III states the problem formulation and detailed description of the NSBPS scheduler. Extensive system level simulation results are introduced in Section IV, and paper is concluded in Section V.

II. System Model

We consider a 5G-NR downlink (DL) MU-MIMO system where there are C cells, each equipped with N_t transmit antennas, and K uniformly distributed user equipment's (UEs) per cell, each with M_r receive antennas. Users are multiplexed by the orthogonal frequency division multiple access (OFDMA). Two types of DL traffic are under assessment as: (a) URLLC bursty FTP3 traffic model with a finite B−byte payload and Poisson arrival process λ, and (b) eMBB full buffer traffic with infinite payload size. The total number of UEs per cell is: $K_{\text{mbb}}+K_{\text{llc}}=K$, where K_{mbb} and K_{llc} are the average numbers of eMBB and URLLC UEs per cell, respectively.

The agile 5G-NR frame structure is adopted [5], where the URLLC and eMBB UEs are scheduled with variable TTI periodicity. As depicted in Fig. 1, eMBB traffic is scheduled with a long TTI of 14-OFDM symbols for SE maximization while URLLC traffic with a shorter TTI of 2-OFDM symbols due to its latency budget. In the frequency domain, the smallest scheduling unit is the physical resource block (PRB), which is 12 sub-carriers and with 15 kHz sub-carrier spacing.

A maximal subset of MU co-scheduled URLLC-eMBB user pairs $\mathtt{G}_c\in\mathcal{K}_c$ is allowed over an arbitrary PRB in the c^{th} cell, where $G_c=\mathbf{card}(\mathtt{G}_c)$, $G_c\le N_t$ is the number of co-scheduled UEs and $\mathcal{K}_c$ is the set of all active UEs in the c^{th} cell. Since $N_t\le KM_r$, user selection on top of equal power allocation is assumed for MU pairing. The received DL signal at the k^{th} user from the c^{th} cell can be modeled as

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

where $\mathbf{H}^{\kappa}_{k,c}\in\mathcal{C}^{M_r\times N_t},\forall k\in\{1,\dots,K\},\forall c\in\{1,\dots,C\}$ is the wireless channel observed 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, assuming a single layer transmission per user, where it is given as: $\mathbf{v}^{\kappa}_{k,c}=\left(\mathbf{H}^{\kappa}_{k,c}\right)^{\mathrm{H}}\left(\mathbf{H}^{\kappa}_{k,c}\left(\mathbf{H}^{\kappa}_{k,c}\right)^{\mathrm{H}}\right)^{-1}$. $s^{\kappa}_{k,c}$ and $\mathbf{n}^{\kappa}_{k,c}$ denote the transmitted symbol and the additive white Gaussian noise at the k^{th} user, respectively. The first and second summation terms represent the intra-cell inter-user and inter-cell interference, generated from either the URLLC or eMBB traffic. In this work, the 3GPP 3D spatial channel model [9] is adopted, where the DL channel coefficient observed by the m^{th} receive antenna from the n^{th} transmit antenna is composed from Q spatial clusters, each with Z rays as

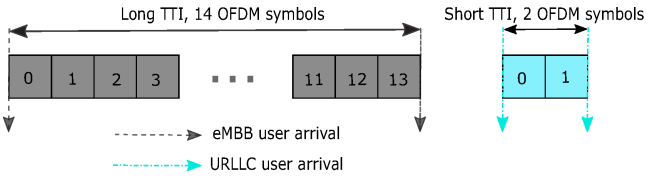

Fig. 1. Flexible TTI scheduling in 5G NR.

$$h^{\kappa}_{(m,n)_k}=\frac{1}{\sqrt{Q}}\sum_{q=0}^{Q-1}\sqrt{\delta_k}\,\mathcal{G}_{q,k}\,r_{(m,n,q)_k},\tag{2}$$

where $\delta_k=\ell\epsilon_k^{\varrho}\mu_k$, ℓ and μ_k are the propagation and shadow fading coefficients, respectively, and ϵ_k^{ϱ} is the distance, with ϱ as the pathloss factor, and $\mathcal{G}_{q,k}\sim\mathbb{CN}(0,1)$. The steering factor $r_{(m,n,q)_k}$ is given by

$$r_{(m,n,q)_k}=$$
$$\sqrt{\frac{\xi\psi}{Z}}\sum_{z=0}^{Z-1}\begin{pmatrix}\sqrt{\mathfrak{D}_{\text{BS}}^{m,n,q,z}(\theta_{\text{AoD}},\varphi_{\text{EoD}})}\,e^{j(\eta d\overline{f}+\Phi_{m,n,q,z})}\\ \times\sqrt{\mathfrak{D}_{\text{UE}}^{m,n,q,z}(\theta_{\text{AoA}},\varphi_{\text{EoA}})}\,e^{j(\eta d\sin(\theta_{\text{m,n,q,z,AoA}}))}\\ \times e^{j\eta||s||\cos(\varphi_{\text{m,n,q,z,EoA}})\cos(\theta_{\text{m,n,q,z,AoA}}-\theta_s)t}\end{pmatrix},\tag{3}$$

where ξ and ψ are the power and large-scale coefficients, $\mathfrak{D}_{\text{BS}}$ and $\mathfrak{D}_{\text{UE}}$ are the antenna patterns at the BS and UE, respectively, η is the wave number, θ denotes the horizontal angle of arrival θ_{AoA} and departure θ_{AoD}, while φ denotes the elevation angle of arrival φ_{EoA} and departure φ_{EoD}, respectively. s is the speed of the k^{th} user, $\overline{f}=f_x\cos\theta_{\text{AoD}}\cos\varphi_{\text{EoD}}$ is the displacement vector of the uniform linear transmit array.

The received signal at the k^{th} user is decoded by applying the antenna combining as: $\left(\mathrm{y}^{\kappa}_{k,c}\right)^{*}=\left(\mathbf{u}^{\kappa}_{k,c}\right)^{\mathrm{H}}\mathrm{y}^{\kappa}_{k,c}$, where $\mathbf{u}^{\kappa}_{k,c}$ is designed by the linear minimum mean square error interference rejection combining (LMMSE-IRC) receiver [10]. The received SINR level at the k^{th} user is then calculated as

$$\Upsilon^{\kappa}_{k,c}=\frac{p^c_k\left\|\mathbf{H}^{\kappa}_{k,c}\mathbf{v}^{\kappa}_{k,c}\right\|^2}{1+\sum\limits_{g\in\mathtt{G}_c,g\neq k}p^c_g\left\|\mathbf{H}^{\kappa}_{k,c}\mathbf{v}^{\kappa}_{g,c}\right\|^2+\sum\limits_{j\in C,j\neq c}\sum\limits_{g\in\mathtt{G}_j}p^j_g\left\|\mathbf{H}^{\kappa}_{g,j}\mathbf{v}^{\kappa}_{g,j}\right\|^2},\tag{4}$$

where p^c_k is the transmit power intended for the k^{th} user. Then, the k^{th} user received rate on a given PRB is given by

$$r^{\kappa}_{k,rb}=\log_2\left(1+\frac{1}{G_{k,c}}\Upsilon^{\kappa}_{k,c}\right).\tag{5}$$

Accordingly, the user SINR levels across different $\mathcal{N}$ sub-carriers are mapped into a single effective SINR using the effective exponential SNR mapping [11] as

$$\left(\Upsilon^{\kappa}_{k,c}\right)^{\text{eff.}}=-\partial\ln\left(\frac{1}{\mathcal{N}}\sum_{i=1}^{\mathcal{N}}e^{-\frac{\left(\Upsilon^{\kappa}_{k,c}\right)^i}{\partial}}\right),\tag{6}$$

with ∂ as a calibration parameter.

III. Proposed NSBPS Scheduler

A. Problem Formulation

Under a 5G-NR system, there are user-centric, instead of network-centric, QoS utility functions. These are highly coupled and need to be reliably fulfilled, e.g., eMBB rate maximization and URLLC latency minimization as

$$\forall k_{\text{mbb}} \in \mathcal{K}_{\text{mbb}} : R_{\text{mbb}} = \underset{k_{\text{mbb}} \in \mathcal{K}_{\text{mbb}}}{\arg\max} \sum_{k_{\text{mbb}}=1}^{K_{\text{mbb}}} \sum_{rb \in \Xi^{\text{mbb}}_{k_{\text{mbb}}}} \beta_{k_{\text{mbb}}} r^{\text{mbb}}_{k_{\text{mbb}},rb}, \quad (7)$$

$$\forall k_{\text{llc}} \in \mathcal{K}_{\text{llc}} : \underset{k_{\text{llc}} \in \mathcal{K}_{\text{llc}}}{\arg\min} \left(\Psi_{k_{\text{llc}}} \right), \Psi_{k_{\text{llc}}} \leq 1\,\text{ms}, \quad (8)$$

where R_{mbb} is the overall eMBB ergodic capacity, $\mathcal{K}_{\text{mbb}}$ and $\mathcal{K}_{\text{llc}}$ represent the active sets of eMBB and URLLC users, respectively, $\Xi^{\text{mbb}}_{k_{\text{mbb}}}$ and $\beta_{k_{\text{mbb}}}$ denote the granted set of PRBs and a priority factor of the k^{th} eMBB user. $\Psi_{k_{\text{llc}}}$ is the URLLC target one-way latency, which is expressed as

$$\Psi_{k_{\text{llc}}} = \Lambda_{\text{q}} + \Lambda_{\text{bsp}} + \Lambda_{\text{fa}} + \Lambda_{\text{tx}} + \Lambda_{\text{uep}}, \quad (9)$$

where $\Lambda_{\text{q}}, \Lambda_{\text{bsp}}, \Lambda_{\text{fa}}, \Lambda_{\text{tx}}, \Lambda_{\text{uep}}$ 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} and Λ_{uep} are bounded by 3-OFDM symbol duration [12], due to the enhanced processing capabilities with the 5G-NR. Hence, Λ_{tx} and Λ_{q} are the major impediment against achieving the hard URLLC latency budget. Λ_{tx} depends on the outage SINR level as given by

$$\Lambda_{\text{tx}} = \frac{\text{B}}{\left(\Xi^{\text{llc}}_{k_{\text{llc}}} \log_2 \left(1 + \frac{\Upsilon^{\text{llc}}_{k_{\text{llc}}}}{\digamma} \right) \right)}, \quad (10)$$

where $\digamma$ is the outage gap between the expected and actual received SINR levels. The URLLC queuing delay Λ_{q} can be modeled by the $\mathcal{A}/\mathcal{A}/\mathsf{a}/\phi$ queuing model [13], where the first $\mathcal{A}$ denotes a Poisson packet arrival, second $\mathcal{A}$ means exponential service times out of the queue, notation a represents the maximum number of the URLLC simultaneous transmissions, and notation ϕ implies that an arriving URLLC packet will be dropped if there are ϕ outstanding packets, worth of more than 1 ms in the queue. Thus, the probability of the URLLC reliability loss, i.e., $\Lambda_{\text{q}} \geq 1$ ms, is given as

$$\rho_{rl} = \left(\rho_0 \frac{\mathsf{a}^{\mathsf{a}}}{\mathsf{a}!} \right) \rho^{\phi}, \quad (11)$$

where ρ_0 is the probability of the queue being empty, and $\rho = \left(\frac{\lambda}{\mathsf{a}\mathcal{O}}\right)$, with $\frac{1}{\mathcal{O}}$ as the mean service time. Thus, to achieve the critical URLLC latency, the transmission and queuing delays should be always minimized to provide further allowance for the re-transmission delay. This can be achieved by guaranteeing a sufficient outage SINR level or allocating excessive PRBs to URLLC traffic in order to further minimize ρ_{rl}. In both cases, the eMBB utility function in (7) will be ill-optimized, leading to a severe degradation of the network SE.

B. Description of The Proposed NSBPS Scheduler

The proposed NSBPS scheduler seeks to simultaneously cross-optimize the joint objectives of the eMBB and URLLC traffic. Thus, the critical URLLC latency deadline is satisfied regardless of the system loading while reaching the best achievable eMBB performance. In the following sub-sections, we describe the proposed NSBPS scheduler in-detail.

At the BS side:

At an arbitrary TTI instance, if there are no newly incoming URLLC packets, NSBPS allocates single-user (SU) dedicated resources to the new/buffered eMBB traffic based on the standard proportional fair (PF) metric as

$$\Theta\left\{\text{PF}_{k_{\text{mbb}}}\right\} = \frac{r^{\text{mbb}}_{k_{\text{mbb}},rb}}{\overline{r}^{\text{mbb}}_{k_{\text{mbb}},rb}}, \quad (12)$$

$$k^{*}_{\text{mbb}} = \underset{k_{\text{mbb}} \in \mathcal{K}_{\text{mbb}}}{\arg\max} \ \Theta\left\{\text{PF}_{k_{\text{mbb}}}\right\}, \quad (13)$$

where $\overline{r}^{\text{mbb}}_{k_{\text{mbb}},rb}$ is the average delivered data rate of the k^{th} user. If sporadic DL URLLC packets arrive at the BS while sufficient radio resources are instantly available, the NSBPS scheduler immediately overpowers the eMBB traffic SU priority and assigns SU resources to incoming URLLC traffic based on the weighted PF (WPF) criteria instead as: $\Theta\left\{\text{WPF}_{k_\kappa}\right\} = \frac{r^{\kappa}_{k,rb}}{\overline{r}^{\kappa}_{k,rb}} \beta_{k_\kappa}$, with $\beta_{k_{\text{llc}}} \gg \beta_{k_{\text{mbb}}}$ for immediate URLLC SU scheduling.

However, under a large offered load, which is envisioned with the 5G-NR, schedulable resources may not be instantly available for critical URLLC traffic and accordingly, significant queuing delays are foreseen. In such case, NSBPS scheduler first attempts to fit the URLLC packets within an active eMBB traffic in a normal and non-biased MU transmission, based on a conservative $\gamma-$orthogonality threshold, where $\gamma \rightarrow [0, 1]$. Thus, the incoming URLLC traffic can only be paired with an active eMBB user if they satisfy:

$$1 - \left| \left(\mathbf{v}^{\text{mbb}}_{k_{\text{mbb}}} \right)^{\text{H}} \mathbf{v}^{\text{llc}}_{k_{\text{llc}}} \right|^2 \geq \gamma. \quad (14)$$

with $\forall k_{\text{mbb}} \in \{1, \ldots, K_{\text{mbb}}\}, \forall k_{\text{llc}} \in \{1, \ldots, K_{\text{llc}}\}$. The conservative orthogonality threshold is enforced to safeguard the URLLC traffic from potential inter-user interference. However, if the spatial degrees of freedom (SDoFs) are limited within a TTI, i.e., system is incapable to jointly process several signals between different transceivers on the same resources, and such orthogonality can not be instantly offered, NSBPS scheduler immediately alters the system optimization objective into a region that satisfies the URLLC outage requirements, while imposing minimal loss to the eMBB performance. Thus, the scheduler enforces an instant, biased and controlled MU transmission between URLLC-eMBB user pair. The URLLC outage is guaranteed by satisfying the following conditions,

$$\textbf{rank}\left\{ \left(\mathbf{u}^{\text{llc}}_{k} \right)^{\text{H}} \mathbf{H}^{\text{llc}}_{k} \mathbf{v}^{\text{llc}}_{k} \right\} \sim \text{full}, \quad (15)$$

$$\textbf{rank}\left\{ \left(\mathbf{u}^{\text{llc}}_{k} \right)^{\text{H}} \mathbf{H}^{\text{llc}}_{k} \left(\mathbf{v}^{\text{mbb}}_{k^{\diamond}} \right)' \right\} \sim 0, \quad (16)$$

where $\left(\mathbf{v}^{\text{mbb}}_{k^{\diamond}} \right)'$ denotes the updated precoder of the co-scheduled eMBB user with the incoming URLLC user. Thus,

an arbitrary discrete Fourier transform spatial subspace $\mathbf{v}_{\text{ref}}(\theta)$, pointing towards angle θ, is constructed by

$$\mathbf{v}_{\text{ref}}(\theta) = \left(\frac{1}{\sqrt{N_t}}\right)\left[1, e^{-j2\pi\Delta\cos\theta}, \ldots, e^{-j2\pi\Delta(N_t-1)\cos\theta}\right]^{\text{T}}, \tag{17}$$

where Δ is the absolute antenna spacing. Next, the NSBPS searches for one active eMBB user whose precoder is closest possible to the reference subspace as

$$k^{\diamond}_{\text{mbb}} = \arg\min_{\mathcal{K}_{\text{mbb}}} \mathbf{d}\left(\mathbf{v}^{\text{mbb}}_{k}, \mathbf{v}_{\text{ref}}\right), \tag{18}$$

with the Euclidean distance between $\mathbf{v}^{\text{mbb}}_{k}$ and $\mathbf{v}_{\text{ref}}$ given by

$$\mathbf{d}\left(\mathbf{v}^{\text{mbb}}_{k}, \mathbf{v}_{\text{ref}}\right) = \frac{1}{\sqrt{2}}\left\|\mathbf{v}^{\text{mbb}}_{k}\left(\mathbf{v}^{\text{mbb}}_{k}\right)^{\text{H}} - \mathbf{v}_{\text{ref}}\mathbf{v}^{\text{H}}_{\text{ref}}\right\|. \tag{19}$$

Then, scheduler instantly projects the precoder vector of the selected eMBB user $\mathbf{v}^{\text{mbb}}_{k^{\diamond}}$ onto $\mathbf{v}_{\text{ref}}$ as given by

$$\left(\mathbf{v}^{\text{mbb}}_{k^{\diamond}}\right)' = \frac{\mathbf{v}^{\text{mbb}}_{k^{\diamond}} \cdot \mathbf{v}_{\text{ref}}}{\left\|\mathbf{v}_{\text{ref}}\right\|^2} \times \mathbf{v}_{\text{ref}}, \tag{20}$$

where $\left(\mathbf{v}^{\text{mbb}}_{k^{\diamond}}\right)'$ is the updated eMBB user precoder. The NSBPS scheduler then instantly schedules the incoming URLLC traffic over shared resources with the impacted eMBB user. Since the instant precoder projection is transparent to the victim eMBB user, it exhibits a SE projection loss. However, eMBB loss is constrained minimum, especially under high eMBB user load, e.g., NSBPS scheduler has a higher probability to find an eMBB user whose precoder is originally aligned within $\mathbf{v}_{\text{ref}}$, such that the instant projection process would not greatly impact its achievable capacity. Finally, the BS acknowledges the URLLC user by a single-bit Boolean co-scheduling indication $\alpha = 1$, to be instantly transmitted in the user-centric control channel.

At the URLLC user side:

Upon reception of $\alpha = 1$, the URLLC user realizes that its granted resources, from the scheduling grant, are shared with an active eMBB user whose transmission is aligned within the reference subspace $\mathbf{v}_{\text{ref}}$. Thus, the first-stage decoder matrix of the URLLC user is constructed by a standard LMMSE-IRC receiver to reject the inter-cell interference as

$$\left(\mathbf{u}^{\text{llc}}_{k}\right)^{(1)} = \left(\mathbf{H}^{\text{llc}}_{k}\mathbf{v}^{\text{llc}}_{k}\left(\mathbf{H}^{\text{llc}}_{k}\mathbf{v}^{\text{llc}}_{k}\right)^{\text{H}} + \mathbf{W}\right)^{-1}\mathbf{H}^{\text{llc}}_{k}\mathbf{v}^{\text{llc}}_{k}, \tag{21}$$

where the interference covariance matrix is given by

$$\mathbf{W} = \mathbb{E}\left\{\mathbf{H}^{\text{llc}}_{k}\mathbf{v}^{\text{llc}}_{k}\left(\mathbf{H}^{\text{llc}}_{k}\mathbf{v}^{\text{llc}}_{k}\right)^{\text{H}}\right\} + \sigma^2\mathrm{I}_{M_r}, \tag{22}$$

where I_{M_r} is $M_r \times M_r$ identity matrix. The IRC vector $\left(\mathbf{u}^{\text{llc}}_{k}\right)^{(1)}$ is then de-oriented to be aligned within one possible null space of the effective inter-user interference subspace $\mathbf{H}^{\text{llc}}_{k}\mathbf{v}_{\text{ref}}$, as expressed by

$$\left(\mathbf{u}^{\text{llc}}_{k}\right)^{(2)} = \left(\mathbf{u}^{\text{llc}}_{k}\right)^{(1)} - \frac{\left(\left(\mathbf{u}^{\text{llc}}_{k}\right)^{(1)} \cdot \mathbf{H}^{\text{llc}}_{k}\mathbf{v}_{\text{ref}}\right)}{\left\|\mathbf{H}^{\text{llc}}_{k}\mathbf{v}_{\text{ref}}\right\|^2} \times \mathbf{H}^{\text{llc}}_{k}\mathbf{v}_{\text{ref}}. \tag{23}$$

This way, the final URLLC decoder vector $\left(\mathbf{u}^{\text{llc}}_{k}\right)^{(2)}$ exhibits no inter-user interference, providing the URLLC user with a robust decoding ability.

C. Analytic Analysis Compared to State of The Art

We compare the performance of the proposed NSBPS scheduler against the state-of-the-art schedulers as follows:

1. Punctured scheduler (PS) [8]: the URLLC traffic is always assigned a higher scheduling priority. If radio resources are not available, PS scheduler instantly overwrites part of the ongoing eMBB transmissions, i.e., immediately stop an ongoing eMBB transmission, for instant URLLC scheduling. PS scheduler shows significant improvement of the URLLC latency performance at the expense of highly degraded SE.

2. Multi-user punctured scheduler (MUPS) [14]: in our past work, we considered a MU scheduler on top of the PS scheduler. MUPS first attempts to achieve a successful MU-MIMO transmission between a URLLC-eMBB user pair; however, it is a transparent, non-biased and non-controlled MU-MIMO. If the SDoFs are limited, MUPS scheduler rolls back to PS scheduler. MUPS has shown an improved performance tradeoff between system SE and URLLC latency; however, with a limited and non-robust gain, due to the non-controlled MU-MIMO and the SE-less efficient PS events.

Accordingly, the aggregate eMBB user rate can be linearly calculated from the individual sub-carrier rates for simplicity, assuming OFDMA flat fading channels, as

$$r^{\text{mbb}}_{k_{\text{mbb}}} = \Xi^{\text{mbb}}_{k_{\text{mbb}}} r^{\text{mbb}}_{k_{\text{mbb}},rb}. \tag{24}$$

Then, the portion of the radio resources $\mathit{\Gamma}^{\text{llc}}_{k_{\text{mbb}}}$ allocated to the k^{th} eMBB user, and being altered by the sporadic URLLC traffic, can be expressed by a set of random variables, as

$$\mathbf{\Gamma} = \left(\mathit{\Gamma}^{\text{llc}}_{k_{\text{mbb}}} \mid k_{\text{mbb}} \in \mathcal{K}_{\text{mbb}}\right). \tag{25}$$

Since URLLC packets are of small payload size, it is reasonably to assume that $\mathit{\Gamma}^{\text{llc}}_{k_{\text{mbb}}} \leq \Xi^{\text{mbb}}_{k_{\text{mbb}}}$ is almost surely satisfied. Hence, the actual eMBB rate is formulated by the joint URLLC-eMBB rate allocation function, given by

$$R_{k_{\text{mbb}}} = \mathcal{F}\left(\Xi^{\text{mbb}}_{k_{\text{mbb}}}, \mathit{\Gamma}^{\text{llc}}_{k_{\text{mbb}}}\right). \tag{26}$$

For an instance, if an eMBB user is allocated SU dedicated resources, then $\mathcal{F}\left(\Xi^{\text{mbb}}_{k_{\text{mbb}}}, \mathit{\Gamma}^{\text{llc}}_{k_{\text{mbb}}}\right) = \Xi^{\text{mbb}}_{k_{\text{mbb}}} r^{\text{mbb}}_{k_{\text{mbb}},rb}$. with no capacity loss. However, due to the prioritized URLLC traffic, the actual eMBB user rate suffers a loss over a portion of the allocated resources, expressed by the rate loss function Π as

$$\mathcal{F}\left(\Xi^{\text{mbb}}_{k_{\text{mbb}}}, \mathit{\Gamma}^{\text{llc}}_{k_{\text{mbb}}}\right) = \Xi^{\text{mbb}}_{k_{\text{mbb}}} r^{\text{mbb}}_{k_{\text{mbb}},rb}\left(1 - \Pi\right), \tag{27}$$

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

$$\mathcal{Q}^{\text{mbb}}_{k} = \frac{1}{\left[\left(\mathbf{H}^{\text{mbb}}_{k}\left(\mathbf{v}^{\text{mbb}}_{k^{\diamond}}\right)'\right) \times \left(\mathbf{H}^{\text{mbb}}_{k}\left(\mathbf{v}^{\text{mbb}}_{k^{\diamond}}\right)'\right)^{\text{H}}\right]^{-1}}, \tag{28}$$

where $\mathcal{Q}^{\text{mbb}}_{k}$ is the achievable post-projection channel gain of the k^{th} eMBB user, and its magnitude can be rewritten in terms of the precoder projection loss, i.e., the *on-the-fly* eMBB precoder update from $\mathbf{v}^{\text{mbb}}_{k^{\diamond}}$ to $\left(\mathbf{v}^{\text{mbb}}_{k^{\diamond}}\right)'$, as

$$\mathcal{Q}_k^{\text{mbb}} = \left\| \mathbf{H}_k^{\text{mbb}} \mathbf{v}_{k^\diamond}^{\text{mbb}} \right\|^2 \times \sin^2\left(\theta_{[\mathbf{v}_{k^\diamond}^{\text{mbb}}, (\mathbf{v}_{k^\diamond}^{\text{mbb}})']}\right), \tag{29}$$

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

$$\overset{\text{NSBPS}}{\Pi} = \left(\frac{\Gamma_{k_{\text{mbb}}}^{\text{llc}}}{\Xi_{k_{\text{mbb}}}^{\text{mbb}}}\right) \times \sin^2\left(\theta_{[\mathbf{v}_{k^\diamond}^{\text{mbb}}, (\mathbf{v}_{k^\diamond}^{\text{mbb}})']}\right). \tag{30}$$

Due to the constraints in (14) and (18), the projection loss is always guaranteed to be minimized, i.e., $\sin^2\left(\theta_{[\mathbf{v}_{k^\diamond}^{\text{mbb}}, (\mathbf{v}_{k^\diamond}^{\text{mbb}})']}\right) \ll 1$. For the PS scheduler, the rate loss function is expressed in terms of the entire URLLC resources inducing the resource allocation of the eMBB user, since the eMBB transmission is instantly stopped over these resources, as

$$\overset{\text{PS}}{\Pi} = \left(\frac{\Gamma_{k_{\text{mbb}}}^{\text{llc}}}{\Xi_{k_{\text{mbb}}}^{\text{mbb}}}\right). \tag{31}$$

Finally, the MUPS scheduler exhibits an average eMBB capacity loss due to the persistent PS events, if the normal MU-MIMO scheduler fails; thus, the rate loss can be given as

$$\overset{\text{MUPS}}{\Pi} = \Phi\left(\frac{\Gamma_{k_{\text{mbb}}}^{\text{llc}}}{\Xi_{k_{\text{mbb}}}^{\text{mbb}}}\right), \tag{32}$$

where $\Phi \leq 1$ is a fraction to indicate the probability density of rolling back to PS scheduler, under a specific cell loading. Hence, the average eMBB user rate can be calculated as

$$\overline{R}_{k_{\text{mbb}}} = \Xi_{k_{\text{mbb}}}^{\text{mbb}} r_{k_{\text{mbb}},rb}^{\text{mbb}} \left(1 - \mathbb{E}\left\{\Pi\right\}\right). \tag{33}$$

Based on (26) - (33), it can be concluded that the proposed NSBPS scheduler provides the best achievable eMBB and URLLC joint performance against state-of-the-art schedulers.

IV. Simulation Results

In this section, we present the extensive SLS results of the NSBPS scheduler, following the 5G-NR specifications, where the main simulation parameters are listed in Table I.

Fig. 2 shows the URLLC average one-way latency Ψ at the 10^{-5} outage probability, under proposed NSBPS, PS, MUPS, and WPF schedulers. On the top left, a close snap of the complementary cumulative distribution function (CCDF) of the URLLC latency distribution is further presented. We define the cell load setup by: $\Omega = (K_{\text{mbb}}, K_{\text{llc}})$. The proposed NSBPS scheduler clearly provides a significantly robust and steady URLLC latency against different cell load conditions, and hence, independently from the aggregate levels of interference. The overall performance gain of the NSBPS scheduler is due to: 1) the guaranteed instantaneous URLLC scheduling without queuing in a controlled (almost surely occurs), biased (for the sake of the URLLC user), and semi-transparent (URLLC user is aware of it) MU transmission, leading to no inter-user interference at the URLLC user, 2) the constrained-minimum eMBB user rate loss function, and 3) the enforced regularization of the inter-cell interference spatial distribution within a limited span, due to the fixed subspace projection, and hence, the linear MMSE-IRC receiver nulls the average inter-cell interference more efficiently and with improved SDoFs.

The PS scheduler shows an optimized URLLC latency in the low load region, at the expense of degraded eMBB performance. However, in the high load region when the inter-cell interference levels are extreme, PS scheduler provides a degraded URLLC latency performance due to the experienced re-transmissions and degraded capacity per PRB. The MUPS scheduler shows a fair tradeoff between URLLC latency and the eMBB SE, where the non-controlled URLLC-eMBB MU-MIMO transmissions reduce the URLLC decoding ability. Finally, the WPF scheduler exhibits the worst URLLC latency performance, where the URLLC packets are queued for multiple TTIs if the radio resources are not instantly schedulable.

As shown in Fig. 3, the empirical CDF (ECDF) of the average cell throughput in Mbps is presented. The NSBPS scheduler provides the best achievable cell throughput compared to other schedulers, due to the always constrained-

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: 5, 10 and 20 users/cell eMBB: 5 , 10 and 20 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

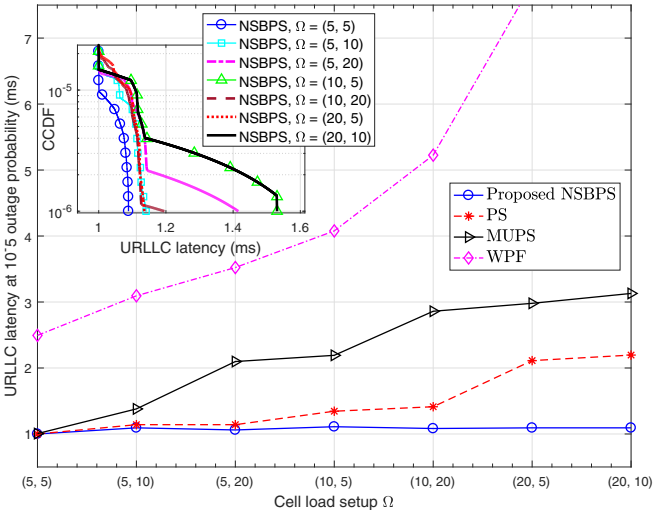

Fig. 2. URLLC one-way latency Ψ at 10^{-5}outage.

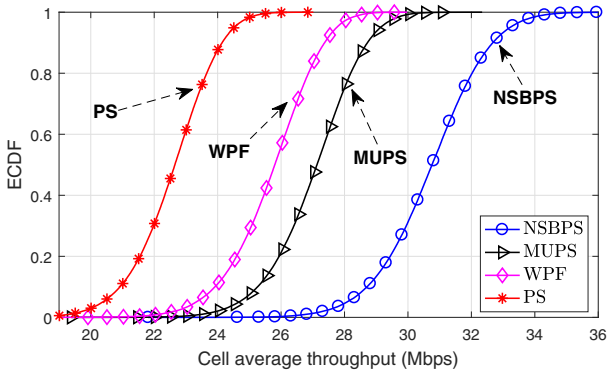

Fig. 3. Cell average throughput for $\Omega = (5, 5)$.

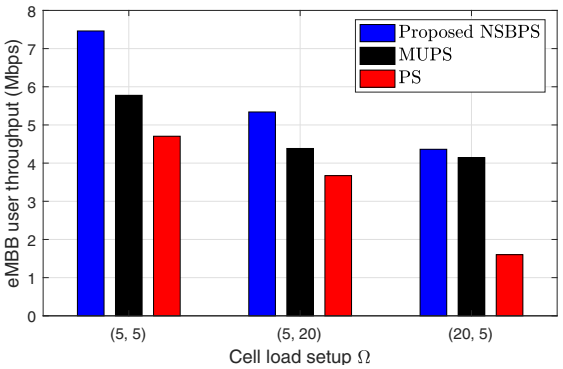

Fig. 4. eMBB average user throughput.

minimum rate loss function of the victim eMBB users. The PS scheduler exhibits severe loss of the network SE due to the punctured eMBB transmissions. However, the WPF scheduler achieves an improved capacity since no puncture-events are allowed; however, at the expense of the worst URLLC latency. Finally, the MUPS scheduler shows further improved capacity, due to the successful MU events; however, with limited MU gain since when a successful MU pairing is not possible, MUPS falls back to SE-less-efficient PS scheduler.

Examining the eMBB performance, Fig. 4. shows the average eMBB user throughput in Mbps, for all schedulers under evaluation, where similar conclusions can be clearly obtained. For instance, with $\Omega = (5, 5)$, where the system SDoFs are limited by the small number of active eMBB users, i.e., $K_{\text{mbb}} = 5$, the proposed NSBPS shows a gain $\sim 28.9\%$ in the eMBB user throughput than the MUPS scheduler. Under such SDoF-limited state, the MUPS scheduler is highly likely to roll back to PS scheduler, i.e., $\Phi \sim 1$, while the NSBPS forcibly enforces these missing SDoFs, sufficient enough to instantly fit the URLLC traffic within an eMBB transmission.

Finally, Table II presents the achievable MU throughput gain of the NSBPS and MUPS schedulers. The best achievable MU gain of the NSBPS over the MUPS scheduler is obtained when the system is originally SDoF-limited, i.e., $\Omega = (5, 5)$. With SDoF-rich loading states such as $\Omega = (20, 5)$, the MUPS scheduler rarely falls back to PS scheduler, i.e., $\Phi \sim 0$, and hence, an improved MU gain is achieved.

Table II
AVERAGE MU GAIN OF THE NSBPS AND MUPS SCHEDULERS.

Scheduler	$\Omega = (5, 5)$	$\Omega = (5, 20)$	$\Omega = (20, 5)$
MUPS (Mbps)	7.69	12.13	23.05
NSBPS (Mbps)	22.92	24.91	27.78
Gain (%)	+198.04	+105.35	+20.52

V. Conclusion

A null space based preemptive scheduler (NSBPS) has been proposed for joint 5G URLLC and eMBB traffic. The proposed NSBPS scheduler aims to fulfill a constraint-coupled objective, for which the URLLC quality of service is almost surely guaranteed while achieving the maximum possible ergodic capacity. Extensive system level simulations and analytic gain analysis have been conducted for performance evaluation. Compared to the state-of-the-art scheduler proposals from academia and industry, the proposed NSBPS shows extreme robustness of the URLLC latency performance, i.e., regardless of the cell loading, and aggregate interference levels, while providing significantly improved eMBB performance. A comprehensive study on the performance of the proposed scheduler will be considered in a future work.

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.

References

[1] NR and NG-RAN overall description; Stage-2 (Release 15), 3GPP, TS 38.300, V2.0.0, Dec. 2017.
[2] Service requirements for the 5G system; Stage-1 (Release 16), 3GPP, TS 22.261, V16.2.0, Dec. 2017.
[3] Study on new radio access technology; Radio access architecture and interfaces (Release 14), 3GPP, TR 38.801, V14.0.0, March 2017.
[4] B. Soret, P. Mogensen, K. I. Pedersen and M. Aguayo-Torres, "Fundamental tradeoffs among reliability, latency and throughput in cellular networks," *in Proc. IEEE Globecom*, Austin, TX, 2014, pp. 1391-1396.
[5] K. I. Pedersen, G. Berardinelli, F. Frederiksen, P. Mogensen and A. Szufarska, "A flexible 5G frame structure design for FDD cases," *IEEE Commun. Mag.*, vol. 54, no. 3, pp. 53-59, March 2016.
[6] 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.
[7] 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, March 2018.
[8] K.I. Pedersen, G. Pocovi, J. Steiner, and S. Khosravirad, "Punctured scheduling for critical low latency data on a shared channel with mobile broadband," *in Proc. IEEE VTC*, Toronto, 2017, pp. 1-6.
[9] Study on 3D channel model for LTE; Release 12, 3GPP, TR 36.873, V12.7.0, Dec. 2014
[10] 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
[11] S. N. Donthi and N. B. Mehta, "An accurate model for EESM and its application to analysis of CQI feedback schemes," *IEEE Trans. Wireless Commun.*, vol. 10, no. 10, pp. 3436-3448, Oct. 2011.
[12] Physical layer procedures; Evolved universal terrestrial radio access (Release 15), 3GPP, TS 36.213, V15.1.0, March. 2018.
[13] Bertsekas, D. and Gallager, R. (1992). *Data Networks*. 2nd ed. Michigan: Prentice Hall.
[14] 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.