Timely Scheduling of URLLC Packets Using Precoder Compatibility Estimates

Nassar Ksairi* and Marios Kountouris†
*Mathematical and Algorithmic Sciences Lab, Paris Research Center
Huawei Technologies France, 92100 Boulogne-Billancourt, France
†Communication Systems Department, EURECOM, Sophia-Antipolis, France
Emails: nassar.ksairi@huawei.com, marios.kountouris@eurecom.fr

***Abstract*—5G systems need to efficiently support both enhanced mobile broadband traffic (eMBB) and ultra-reliable, low-latency communications (URLLC) traffic. In this paper, the problem of joint eMBB and URLLC scheduling is studied with the objective of supporting the URLLC requirements with minimal impact on eMBB performance. We propose a preemptive scheduler exploiting a precoder compatibility estimate, which measures the degree of similarity between different multiple-input multiple-output (MIMO) channels. This metric allows us to determine which eMBB transmission to pause for serving a URLLC user without recomputing the precoding matrix, thus reducing the processing time needed for multiuser MIMO precoder computation, while both relaxing cell pilot periodicity and reducing URLLC demodulation pilot overhead. This is achieved thanks to the fact that precoder compatibility metrics can be acquired at the user side in an opportunistic manner based on pilot symbols destined to ongoing downlink transmissions. Simulation results assess the performance gains of the proposed scheduler in terms of both URLLC block error rate and eMBB throughput.**

I. Introduction

Ultra-reliable, low-latency communications (URLLC) require extremely low latency (in the order of one to few milliseconds) with very high reliability (99.999%). To support the stringent delay requirements, URLLC transmissions are structured on the basis of very short periods of time, say one to two orthogonal frequency division multiplexing (OFDM) symbols, referred to as *m*ini-slots that are shorter than typical transmission time interval (TTI) values. Since efficient utilization of spectrum dictates coexistence between URLLC and enhanced mobile broadband (eMBB), at least on a portion of the system bandwidth, the packet scheduler should be able to handle heterogeneous traffic and to operate at different TTI values. This leads to a very challenging scheduling problem [1]. For instance, upon their arrival - even during eMBB transmissions that occupy all the time-frequency resources of the coexistence region- URLLC packets should be transmitted in the immediate next mini-slot [2]. One way to perform that is “resource sharing” [2], [3] in which the available degrees of freedom in the spatial and/or the power domains are used to support both the ongoing data transmissions and the incoming URLLC packet. Another way is “preemptive” scheduling (also called puncturing) [2], [4], [5]. It consists in preempting, i.e., pausing, one or more ongoing large-packet transmissions during one or few mini-slots upon short-packet arrivals. Note that in both resource sharing and preemptive scheduling, the multiuser multiple MIMO (MU-MIMO) precoder should in principle be recalculated for the new set of scheduled users during the mini-slot used for URLLC transmission. The time needed for matrix inversions, decompositions, and other costly computations typically involved in MU-MIMO downlink precoding is an important part of the base station (BS) *processing delay* [6] in the end-to-end latency budget. Reducing processing time and complexity is of cardinal importance for meeting URLLC latency requirements. A solution for reducing this delay is proposed in [7]. However, this solution is only applicable to optimal beamforming and not to practically relevant linear precoding schemes. Another issue with MU-MIMO precoder updating during URLLC mini-slots is that it typically requires accurate channel state information (CSI) about the wireless link to the URLLC user to be available at the BS. This requirement translates in frequency division duplexing (FDD) systems into the need for high downlink cell reference signals (CRS) transmissions and CSI feedback periodicities.

We propose a preemptive scheduler that does not require the computationally costly recomputation of the MU-MIMO precoder upon a URLLC arrival. The key idea is to determine which ongoing eMBB transmission to pause based on a *precoder compatibility metric* between the eMBB MIMO precoder and the channel to the URLLC user. By “precoder compatibility” we mean any metric, e.g., channel vectors *co-linearity coefficient*, which provides an indication of the performance that can be obtained from using a given precoder on a given channel even if the precoder in question was not originally computed for transmission on that channel. We also propose a method to acquire precoder compatibility values at the URLLC receiver by sensing the signals at the positions of demodulation reference signal (DMRS) symbols destined to currently ongoing downlink transmissions. This *opportunistic* acquisition alleviates the need for very frequent CRS pilot transmissions, otherwise needed to keep up-to-date channel estimates about URLLC user terminals (UTs). Another important implication is the possibility of reducing DMRS overhead for URLLC to levels as low as zero.

II. System Model

We consider downlink transmissions from a BS equipped with M transmit antennas to single-antenna UTs using OFDM with N_{FFT} subcarriers. We focus on a scenario in which both eMBB and URLLC traffic coexist on at least a portion of the resource grid. The coexistence region is composed of N_{RBG} resource block groups (RBGs), each containing T *mini-slots*, each of which is typically composed of one to two OFDM symbols (see Figure 1). Denote by $\mathcal{K}_b^{\mathrm{eMBB}}$ the set of UTs for which eMBB data packets are spatially multiplexed on the b-th RBG, $b \in \{1, \ldots, N_{\mathrm{RBG}}\}$. For presentation simplicity, we assume that each transmission uses one spatial layer and that $\forall b \in \{1, \ldots, N_{\mathrm{RBG}}\}$, $\mathcal{K}_b^{\mathrm{eMBB}} \neq \emptyset$. In the upper part of Figure 1, blocks of the same color represent eMBB packets that are spatially multiplexed on the same RBG (that is shown in the bottom part of the figure using the same color as the packets occupying it).

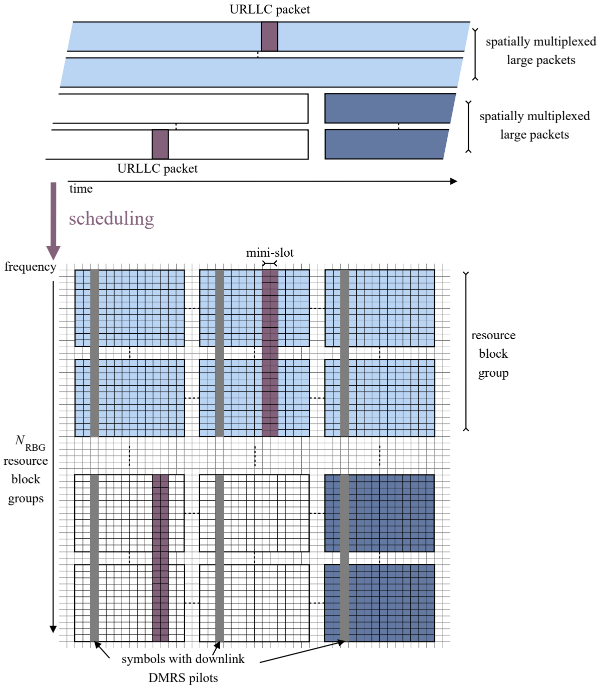
URLLC packet
spatially multiplexed large packets
spatially multiplexed large packets
URLLC packet
time
scheduling
mini-slot
frequency
resource block group
N_{RBG} resource block groups
symbols with downlink DMRS pilots

Fig. 1. Preemptive scheduling for URLLC packets

Let $\mathcal{K}_t^{\mathrm{URLLC}}$ designate the set of UTs for which a URLLC packet arrives at the BS during the t-th mini-slot. We assume that each URLLC transmission occupies one RBG in the frequency domain and one mini-slot in the time domain (as shown in Figure 1). Extension to URLLC transmissions occupying several RBGs, e.g., to improve reliability, is left for future work. The performance measure of interest for eMBB transmissions is the sum spectral efficiency, while for URLLC traffic is block error rate (BLER) within a certain delay. When the modulation and coding scheme (MCS) is fixed, a target BLER can be translated into a target signal-to-interference-plus-noise ratio (SINR) value $\mathrm{SINR}^{\mathrm{tr}}$.

A. Signal Model

We assume that the channel response is constant over the resource elements (REs) of each RBG; this assumption is compatible with the fact that RBGs are the units of time-frequency resources for MIMO precoder assignment in both our system model and in real-world cellular systems. The frequency domain channel coefficient between the m-th BS antenna and k-th UT at any subcarrier within RBG b is denoted by $H_{k,m,b}$. We let $x_{k,t,n}$ represent the unit-variance data symbol transmitted on the n-th subcarrier ($n \in \{0, \ldots, N_{\mathrm{FFT}} - 1\}$) during the t-th mini-slot and we define $\mathbf{H}_{k,b} \triangleq \left[H_{k,1,b} \cdots H_{k,M,b}\right]^{\mathrm{T}}$. While we do not restrict vectors $\mathbf{H}_{k,b}$ to follow specific channel models, the simulation results in Section IV are obtained with realizations of $\mathbf{H}_{k,b}$ generated using a 'sectorized' version of the so-called physical channel model [8]: each user's channel is reducible to $1 \leq P \leq M$ dimensions (or angular bins) covering one of $N_c \geq 1$ ranges of angles of departure (AoD) values that partition the total range $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$. More precisely, denote by $c_k \in \{1, \ldots, N_c\}$ the index of the AoD sector to which the channel to UT k belongs, by ϕ_{c_k} the smallest AoD in that sector and by $\phi_{c_k,p} \triangleq \phi_{c_k} + (p-1)\Delta_\phi$ (for some $0 < \Delta_\phi < \frac{\pi}{2P}$) the AoD associated with the p-th angular bin of sector c_k. Then $\mathbf{H}_{k,b} = \sum_{p=1}^{P} h_{k,b,p}\mathbf{a}\left(\phi_{c_k,p}\right)$ where $h_{k,b,p} \sim \mathcal{CN}\left(0, \frac{1}{\sqrt{P}}\right)$ is the complex channel gain along the p-th angular bin and $\mathbf{a}(\phi) \triangleq \frac{1}{\sqrt{M}}\left[1 \quad e^{-\imath\pi\cos(\phi)} \quad \cdots \quad e^{-\imath\pi(M-1)\cos(\phi)}\right]^{\mathrm{T}}$. Denote by $b_n \in \{1, \ldots, N_{\mathrm{RBG}}\}$ the index of the RBG to which subcarrier n belongs. The channel sample at receiver $k \in \mathcal{K}_{b_n,t}^{\mathrm{URLLC}}$ at RE (t, n) is given by

$$y_{k,t,n} = \sqrt{g_k}\mathbf{w}_{k,b_n}^{\mathrm{T}}\mathbf{H}_{k,b_n}x_{k,t,n} + \sum_{i\in\mathcal{K}_{b_n}^{\mathrm{eMBB}}\cup\mathcal{K}_t^{\mathrm{URLLC}}\setminus\{k\}} \sqrt{g_k}\mathbf{w}_{i,b_n}^{\mathrm{T}}\mathbf{H}_{k,b_n}x_{i,t,n} + z_{k,t,n}, \tag{1}$$

where $\mathbf{w}_{k,b_n} \in \mathbb{C}^{M\times 1}$ is the precoding vector used for transmission on RBG b_n, $z_{k,t,n} \sim \mathcal{CN}(0, \sigma_k^2)$ is the noise sample and g_k is the large-scale fading attenuation between the BS and the k-th UT. The precoding vectors should satisfy the following transmit power constraint

$$\sum_{i\in\mathcal{K}_b^{\mathrm{eMBB}}} \|\mathbf{w}_{i,b}\|^2 \leq P_{\mathrm{BS}}, \quad \forall b \in \{1, \ldots, N_{\mathrm{RB}}\}. \tag{2}$$

To help each UT i in estimating the *effective channel* coefficients $\mathbf{w}_{i,b_n}^{\mathrm{T}}\mathbf{H}_{i,b}$, N_p and N_p^{mini} REs are reserved for DMRS pilot transmission within each RB with eMBB data traffic and mini-slot with URLLC data, respectively. In the next section, we present a scheduling method for URLLC traffic that reduces the DMRS overhead in URLLC mini-slots to as low as $N_p^{\mathrm{mini}} = 0$.

III. Proposed URLLC Scheduler using Precoder Compatibility Metrics

Henceforth, we assume that at most one URLLC data packet may arrive in the BS transmission queue during any mini-slot, i.e., $K_t^{\mathrm{URLLC}} \leq 1$. Once a URLLC packet destined to UT k arrives at the BS, it is scheduled for transmission on the first mini-slot t directly following its arrival (as in the example shown in Figure 1) resulting in $K_t^{\mathrm{URLLC}} = 1$ and $\mathcal{K}_t^{\mathrm{URLLC}} = \{k\}$. This is done (i) by pausing one of the ongoing eMBB transmissions for the duration of that mini-slot on *one* of the N_{RBG} RBGs of the coexistence frequency sub-band (preemptive scheduling) and (ii) *by precoding the URLLC packet using the MIMO precoder of the paused eMBB transmission*. The method we propose to determine the index of the particular RBG to be occupied by the URLLC packet and the particular eMBB transmission to be paused (preempted) is explained in what follows.

A. Precoder Compatibility Metrics

Denote by $H_{k,i,b}^{\mathrm{eff}}$ the effective channel to UT k resulting from the sole use of the MIMO precoder destined to UT j, i.e., $H_{k,i,b}^{\mathrm{eff}} \triangleq \mathbf{w}_{i,b}^{\mathrm{T}} \mathbf{H}_{k,b}$. We introduce a *p*recoder compatibility metric between URLLC UT k and eMBB UT i on RBG b, denoted as $G_{k,j,b}$, as the gain of the above effective channel

$$G_{k,i,b} \triangleq \left| H_{k,i,b}^{\mathrm{eff}} \right|^2 . \tag{3}$$

The proposed precoder compatibility metric measures, roughly speaking, the similarity between the channel of the URLLC user and that of the eMBB user for which the precoder vector $\mathbf{w}_{i,b}$ was originally computed. The rationale is that during preemption, the URLLC packet replaces the eMBB with the most co-linear precoding vector (thus similar channel).

If the co-linearity between the URLLC and eMBB users' channels is high, the former can be served with the precoding vector corresponding to the latter. This significantly reduces the complexity and the latency involved in precoder computation. Now, it might be necessary during some mini-slots to preempt more than one eMBB transmission to guarantee a higher SINR value for the URLLC packet. One reason for occasionally doing so is the need to compensate for any relatively low values of the precoder compatibility metric, which could arise in situations where no eMBB channel vectors are aligned closely enough with the channel vector of the URLLC user. Denote by $\mathcal{J} \subset \mathcal{K}_b^{\mathrm{eMBB}}$ the set of such eMBB UTs with preempted transmissions, by $j \in \mathcal{J}$ the index of the preempted eMBB UT that will 'lend' its MIMO precoder to the URLLC UT k and by $\alpha(j, \mathcal{J}) \triangleq \frac{\sum_{i \in \mathcal{J}} \|\mathbf{w}_{i,b}\|^2}{\|\mathbf{w}_{j,b}\|^2} \geq 1$ the transmit power *boosting factor* that can be applied to the signal transmitted to UT k provided that $|\mathcal{J}| > 1$. Indeed, if the latter condition is satisfied, more power can be assigned to the URLLC packet (than the power $\|\mathbf{w}_{j,b}\|^2$ originally assigned to the paused eMBB transmission) while still satisfying the constraint in (2). This holds because other eMBB transmissions, i.e., the transmissions to the UTs in $\mathcal{J} \setminus \{j\}$, are being paused in this case without being replaced by URLLC transmissions.

B. Problem Formulation and Algorithm Description

The SINR that would result from preempting all transmissions to UTs in $\mathcal{J}$ while using the MIMO precoder originally designed for UT $j \in \mathcal{J}$ to transmit the URLLC packet destined to UT k on RB $b \in \{1, \ldots, N_{\mathrm{RB}}\}$ during mini-slot t is denoted as $\mathrm{SINR}_{k,b,t}(j, \mathcal{J})$ and is given by

$$\mathrm{SINR}_{k,b,t}(j, \mathcal{J}) \triangleq \frac{\alpha(j, \mathcal{J})\, g_k G_{k,j,b}}{\sigma_k^2 + \sum_{i \in \mathcal{K}_b^{\mathrm{eMBB}} \setminus \mathcal{J}} g_k G_{k,i,b}} . \tag{4}$$

We now define $\hat{\mathrm{SINR}}_{k,b,t} \triangleq \frac{\alpha(j,\mathcal{J}) g_k \hat{G}_{k,j,b}}{\sigma_k^2 + \sum_{i \in \mathcal{K}_b^{\mathrm{eMBB}} \setminus \mathcal{J}} g_k \hat{G}_{k,i,b}}$ as an estimate of $\mathrm{SINR}_{k,b,t}$ that is a function of $\left\{ \hat{G}_{k,i,b} \right\}_{i \in \mathcal{K}_b^{\mathrm{eMBB}}, b \in \{1, \ldots, N_{\mathrm{RBG}}\}}$, where $\hat{G}_{k,i,b}$ is a pilot based estimate of $G_{k,i,b}$. We show in the next subsection that $\hat{G}_{k,i,b}$ can be obtained at the URLLC receiver side by opportunistically relying on the DMRS pilots that are destined to the UTs in $\mathcal{K}_b^{\mathrm{eMBB}}$. While the thus defined $\hat{\mathrm{SINR}}_{k,b,t}$ is not optimal with respect to a known estimation optimality criterion, it is clear that its value will be close to the actual SINR value at least when the estimation error variance associated with $\hat{G}_{k,i,b}$ is small enough, e.g., when the number of pilot symbols used in obtaining these effective channel estimates is large enough. The above definition of $\hat{\mathrm{SINR}}_{k,b,t}$ also has the advantage of low computational complexity and of merely relying on the channel estimation modules that are already present in all UTs. Furthermore, good performance results are obtained based on these SINR estimates under realistic simulations settings as validated in Section IV.

The set of eMBB UTs, denoted $\mathcal{J}_{k,b,t}$, whose ongoing data transmissions should be paused and the one among them, denoted by $j_{k,b,t}$, which 'lends' its precoder to the URLLC transmission, is determined by solving

$$\begin{aligned} &\{\mathcal{J}_{k,b,t}, j_{k,b,t}\} = \\ &\qquad \operatorname{argmin}_{\mathcal{J} \subset \mathcal{K}_b^{\mathrm{eMBB}}; j \in \mathcal{J}; \hat{\mathrm{SINR}}_{k,b,t}(j,\mathcal{J}) \geq \mathrm{SINR}^{\mathrm{tr}}} |\mathcal{J}| . \end{aligned} \tag{5}$$

When the above optimization problem is to be solved at the BS side, the precoder compatibility metrics $\left\{ \hat{G}_{k,i,b} \right\}_{i \in \mathcal{K}_b^{\mathrm{eMBB}}, b \in \{1, \ldots, N_{\mathrm{RBG}}\}}$ need to be fed back by the URLLC UT. Otherwise, only the outcome of the optimization, i.e., $\mathcal{J}_{k,b,t}$ and $j_{k,b,t}$, needs to be reported.

The constrained minimization carried out in (5) guarantees that the sum eMBB throughput is reduced the least possible due to preemptions while the incoming URLLC packet is very likely to be decoded reliably and in a timely manner. Note that in the proposed scheme, at least one preemption is needed, i.e., ideally $\mathcal{J}_{k,b,t} = \{j_{k,b,t}\}$. Due to the target SINR (reliability) constraint, the problem in (5) is not necessarily feasible. If it is infeasible, we conventionally set $\mathcal{J}_{k,b,t} = \mathcal{K}_b^{\mathrm{eMBB}}$ and $j_{k,b,t} = \arg\max_{j\in\mathcal{K}_b^{\mathrm{eMBB}}} \widehat{\mathrm{SINR}}_{k,b,t}\left(j,\mathcal{K}_b^{\mathrm{eMBB}}\right) = \arg\max_{j\in\mathcal{K}_b^{\mathrm{eMBB}}} \hat{G}_{k,j,b}$. The following proposition provides a *sufficient* feasibility condition.

Proposition 1. *A sufficient condition for the problem in* (5) *to be feasible at least for one value* $b \in \{1,...,N_{\mathrm{RBG}}\}$ *is*

$$\exists j_0 \in \bigcup\nolimits_{b=1\ldots N_{\mathrm{RBG}}} \mathcal{K}_b^{\mathrm{eMBB}}, \hat{G}_{k,j_0,b} \geq \frac{\mathrm{SINR}^{\mathrm{tr}}\left\|\mathbf{w}_{j_0,b}\right\|^2}{g_k P_{\mathrm{BS}}/\sigma_k^2}. \tag{6}$$

Proof. A sufficient condition for feasibility can be obtained by assuming that the power constraint in (2) is met with equality and by preempting *all* ongoing eMBB transmissions, i.e., by setting $\mathcal{J}_{k,b,t} = \mathcal{K}_b^{\mathrm{eMBB}}$. The condition in (6) directly follows from referring to (4) and setting $\alpha\left(j_0,\mathcal{J}_{k,b,t}\right) = \left(1/\left\|\mathbf{w}_{j_0,b}\right\|^2\right)\sum_{j\in\mathcal{K}_b^{\mathrm{eMBB}}}\left\|\mathbf{w}_{j,b}\right\|^2 = P_{\mathrm{BS}}/\left\|\mathbf{w}_{j_0,b}\right\|^2$. □

Define for any two nonzero complex valued vectors $\mathbf{a}$ and $\mathbf{b}$ the co-linearity coefficient $\rho\left(\mathbf{a},\mathbf{b}\right) \triangleq \frac{\mathbf{a}^{\mathrm{H}}\mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|}$.

Corollary 1. *Under the assumption of maximum ratio transmission (MRT) beamforming and perfect precoder compatibility estimation, the condition in* (6) *writes as*

$$\exists j_0 \in \bigcup\nolimits_b \mathcal{K}_b^{\mathrm{eMBB}}, \rho\left(\mathbf{H}_{j_0,b},\mathbf{H}_{k,b}\right) \geq \frac{\sigma_k^2\mathrm{SINR}^{\mathrm{tr}}}{g_k P_{\mathrm{BS}}\left\|\mathbf{H}_{k,b}\right\|^2}. \tag{7}$$

Proof. Corollary 1 follows from Proposition 1 by substituting $\mathbf{H}_{j,b}^*/\sqrt{\sum_{i\in\mathcal{K}_b^{\mathrm{eMBB}}}\left\|\mathbf{H}_{i,b}\right\|^2}$ for $\mathbf{w}_{j,b}$ ($j\in\{j_0,k\}$) and inserting $\hat{G}_{k,j_0,b} = \left|\mathbf{w}_{j_0,b}^{\mathrm{T}}\mathbf{H}_{k,b}\right|^2$. □

Corollary 1 implies that the feasibility of the problem in (5) is closely related to the probability of finding a scheduled eMBB user whose channel vector is similar enough to (or has a sufficiently large co-linearity coefficient with respect to) the channel of the URLLC user. Assessing this probability requires taking into account users' density in the cell area, users' channel model and the method used by the 'background' eMBB packet scheduler. While this is out of the scope of this paper, it is clear from (6) and (7) that increasing N_{RBG} and $\sum_{b=1}^{N_{\mathrm{RBG}}} K_b^{\mathrm{eMBB}}$ cannot but increase the probability of a successful URLLC transmission. Finally, due to the term $\left\|\mathbf{H}_{k,b}\right\|^2$ in the right-hand side of (7), the minimum acceptable channel similarity condition is less strict on RBGs with favorable fading for the URLLC UT.

Proposition 2. *When it is feasible, at least one solution to* (5) *is such that* $\mathcal{J}_{k,b,t}$ *is the subset composed of the* $J_{k,b,t} \triangleq \left|\mathcal{J}_{k,b,t}\right|$ *UTs* $j \in \mathcal{K}_b^{\mathrm{eMBB}}$ *with the* $J_{k,b,t}$ *largest values of* $\hat{G}_{k,j,b}$ *and such that* $j_{k,b,t} = \arg\max_{j\in\mathcal{J}_{k,b,t}} \hat{G}_{k,j,b}$.

Proof. Assume that there exists a solution that satisfies $\exists j \in \mathcal{J}_{k,b,t}$ and $\exists i \in \mathcal{K}_b^{\mathrm{eMBB}} \setminus \mathcal{J}_{k,b,t}$ such that $\hat{G}_{k,i,b} > \hat{G}_{k,j,b}$. Define $\tilde{\mathcal{J}}_{k,b,t} \overset{\mathrm{def.}}{=} \{i\} \cup \mathcal{J}_{k,b,t} \setminus \{j\}$. First assume that $j = j_{k,b,t}$. Referring to (4) shows that $\widehat{\mathrm{SINR}}_{k,b,t}\left(i,\tilde{\mathcal{J}}_{k,b,t}\right) > \widehat{\mathrm{SINR}}_{k,b,t}\left(j,\mathcal{J}_{k,b,t}\right) \geq \mathrm{SINR}^{\mathrm{tr}}$. This inequality also holds if $j \in \mathcal{J}_{k,b,t} \setminus \{j_{k,b,t}\}$. Therefore, $\left(i,\tilde{\mathcal{J}}_{k,b,t}\right)$ is a solution to (5), thus proving the first part of the proposition. To prove its second part, take one solution $\left(j_{k,b,t},\mathcal{J}_{k,b,t}\right)$ satisfying the above-mentioned property while $\exists j \in \mathcal{J}_{k,b,t}$ such that $\hat{G}_{k,j,b} > \hat{G}_{k,j_{k,b,t},b}$. Now, note by referring to (4) that $\widehat{\mathrm{SINR}}_{k,b,t}\left(j,\mathcal{J}_{k,b,t}\right) > \widehat{\mathrm{SINR}}_{k,b,t}\left(j_{k,b,t},\mathcal{J}_{k,b,t}\right)$. Thus $\left(j,\mathcal{J}_{k,b,t}\right)$ is also a solution to the problem. Putting all pieces together concludes the proof of Proposition 2. □

Among several solutions associated with the same value of $J_{k,b,t}$, we opt for the one that results in the largest SINR. Finally, actual transmission to UT k takes place on RBG $b_{k,t}^*$

$$b_{k,t}^* \triangleq \arg\max_{b\in\arg\min_b\left|\mathcal{J}_{k,b,t}\right|} \widehat{\mathrm{SINR}}_{k,b,t}\left(j_{k,b,t},\mathcal{J}_{k,b,t}\right) \tag{8}$$

while $\mathcal{J}_{k,b,t} = \emptyset, \forall b \neq b_{k,t}^*$. The problem in (8) is always feasible thanks to the default values we give to $j_{k,b,t}$ and $\mathcal{J}_{k,b,t}$ whenever (5) is infeasible. It can be solved using exhaustive search (N_{RBG} is typically small, say between 20 and 50 in LTE system). The steps involved in the proposed scheme during each mini-slot t with URLLC traffic are summarized in Algorithm 1.

The actual SINR of the transmission to URLLC UT k resulting from applying Algorithm 1 is $\mathrm{SINR}_{k,b_{k,t}^*,t}\left(j_{k,b_{k,t}^*,t},\mathcal{J}_{k,b_{k,t}^*,t}\right)$ where $\mathrm{SINR}_{k,b,t}\left(j,\mathcal{J}\right)$ is the function defined in (4). We now define the associated BLER as $\mathrm{BLER}_k^{\mathrm{URLLC}} \triangleq \Pr\left[\mathrm{SINR}_{k,b_{k,t}^*,t}\left(j_{k,b_{k,t}^*,t},\mathcal{J}_{k,b_{k,t}^*,t}\right) < \mathrm{SINR}^{\mathrm{tr}}\right]$ and the sum eMBB instantaneous spectral efficiency as

$$\begin{aligned} R_t^{\mathrm{eMBB}} &\triangleq \\ &\frac{1}{N_{\mathrm{RBG}}}\sum_{b=1}^{N_{\mathrm{RBG}}}\sum_{j\in\mathcal{K}_b^{\mathrm{eMBB}}}\left(1-\delta_{b,b_{k,t}^*}\mathbb{1}_{\mathcal{J}_{k,b,t}}(j)/T\right)\times \\ &\log_2\left(1+\mathrm{SINR}_{j,b,t}\left(\mathcal{J}_{k,b,t}\right)\right) \end{aligned} \tag{9}$$

where $\mathrm{SINR}_{j,b,t}\left(\mathcal{J}\right) \triangleq \frac{g_j G_{j,j,b}}{\sigma_j^2+\sum_{i\in\mathcal{K}_b^{\mathrm{eMBB}}\setminus\{j\}\cup\mathcal{J}} g_j G_{j,i,b}}$.

Algorithm 1 Precoder Compatibility Based Scheduling

```
Input: $\left\{\hat{G}_{k,j,b}\right\}_{b=1,\ldots,N_{\mathrm{RBG}},k\in\mathcal{K}_b^{\mathrm{eMBB}}}, \sigma_k^2, g_k, \mathrm{SINR}^{\mathrm{tr}}$
for $1 \leq b \leq N_{\mathrm{RBG}}$ do
    $\mathcal{J}_{k,b,t} \leftarrow \emptyset$, $\hat{\mathrm{SINR}}_{k,b,t} \leftarrow -1$
    while $\hat{\mathrm{SINR}}_{k,b,t} < \mathrm{SINR}^{\mathrm{tr}}$ and $|\mathcal{J}_{k,b,t}| < K_b^{\mathrm{eMBB}}$ do
        $j_{k,b,t} \leftarrow \arg\max_{j\in\mathcal{K}_b^{\mathrm{eMBB}}\setminus\mathcal{J}_{k,b,t}} \hat{G}_{k,j,b}$
        $\mathcal{J}_{k,b,t} \leftarrow \mathcal{J}_{k,b,t} \cup \{j_{k,b,t}\}$
        $\hat{\mathrm{SINR}}_{k,b,t} \leftarrow \frac{\alpha(j_{k,b,t},\mathcal{J}_{k,b,t}) g_k \hat{G}_{k,j_{k,b,t},b}}{\sigma_k^2+\sum_{i\in\mathcal{K}_b^{\mathrm{eMBB}}\setminus\mathcal{J}_{k,b,t}} g_k \hat{G}_{k,i,b}}$
    end while
end for
$\mathcal{B}_{k,t}^* \leftarrow \arg\min_{b=1,\ldots,N_{\mathrm{RBG}}} |\mathcal{J}_{k,b,t}|$
$b_{k,t}^* \leftarrow \arg\max_{b\in\mathcal{B}_{k,t}^*} \hat{\mathrm{SINR}}_{k,b,t}\left(j_{k,b,t}, \mathcal{J}_{k,b,t}\right)$
Output: $b_{k,t}^*, \mathcal{J}_{k,b_{k,t}^*,t}, j_{k,b_{k,t}^*,t}$
```

Performance metrics $\mathrm{BLER}_k^{\mathrm{URLLC}}$ and R_t^{eMBB} are numerically evaluated in Section IV.

C. Precoder Compatibility Metric Acquisition

In practice, precoder compatibility can only be estimated based on noisy samples. This can be done, for instance, at the BS side using the CSI computed based on CRS pilots and fed back by the UTs.

More interestingly, $G_{k,j,b}$ can be estimated opportunistically at the URLLC receiver side based on the DMRS pilot destined to eMBB UT $j \in \mathcal{K}_b^{\mathrm{eMBB}}$, in which case $\hat{G}_{k,j,b} \triangleq \left|\hat{H}_{k,j,b}^{\mathrm{eff}}\right|^2$, where $\hat{H}_{k,j,b}^{\mathrm{eff}}$ is the effective channel estimate computed at UT k based on the DMRS pilot symbols originally destined to UT j.

Remark 1. *Another advantage of DMRS based opportunistic acquisition of precoder compatibility metrics (done at the UT side) as opposed to the CRS based counterpart (done at the BS side) is that the channel vector estimates $\hat{\mathbf{H}}_{k,b}$ are typically quantized before being fed back to the BS; this makes them prone to both estimation and quantization noises. Finally, DMRS pilots are present in each TTI, while CRS symbols are typically configured with a larger transmission period, thus resulting in precoder compatibility estimates that are on average more outdated than their DMRS counterparts.*

D. Effect on URLLC Pilot Overhead

Since in the proposed scheme URLLC packets are precoded using the same MIMO precoder vector that was being used for a paused eMBB transmission, the effective channel estimate associated with that precoder is readily available at the URLLC receiver (from the precoder compatibility acquisition step) and can thus be used for coherent demodulation. One can thus in principle set $N_p^{\mathrm{mini}} = 0$, i.e., the URLLC packet can be transmitted without the need to insert any DMRS pilot symbols in it. Denote by $\hat{H}_{k,j_{k,b,t},b}^{\mathrm{eff}}$ the channel estimate thus obtained. Computing $\hat{H}_{k,j_{k,b,t},b}^{\mathrm{eff}}$ cannot be done without knowing the actual symbols that were used as DMRS pilots for UT $j_{k,b,t}$, in addition to their positions. In LTE and 5G systems, these symbols are available to UT k as they are merely a function of the serving cell ID [9] and the current frame number[1]. Based on the signal model in (1), it is easy to show that if least squares (LS) estimation is used, then for all $j \in \mathcal{K}_b$

$$\hat{H}_{k,j,b}^{\mathrm{eff}} = H_{k,j,b}^{\mathrm{eff}} + \tilde{H}_{k,j,b}^{\mathrm{eff}} \qquad (10)$$

where $\tilde{H}_{k,j,b}^{\mathrm{eff}} \sim \mathcal{CN}\left(0, \sigma_k^2/\left(N_P + N_p^{\mathrm{mini}}\right)\right)$.

Example 1. *Sending a quadrature phase shift keying (QPSK) modulated message of 3 bytes using a 2-symbol mini-slot covering 4 RBs requires a code rate of 0.3 if the DMRS overhead is* 30% *and of 0.1875 if $N_p^{\mathrm{mini}} = 0$. On AWGN channels, this translates into a 2-dB reduction of* $\mathrm{SINR}^{\mathrm{tr}}$.

IV. Simulation Results

Simulation results are obtained assuming a pool of 10 eMBB and 10 URLLC UTs whose distances to the BS are uniformly distributed in the interval $[0, 750]$ m. Pathloss coefficients g_k are computed based on these distances using the COST-231 Hata model [10] with a carrier frequency $f_c = 1800$ MHz. Channel vectors $\mathbf{H}_{k,b}$ are generated using the channel model presented in Subsection II-A with $M = 16$, $N_c = 3$ and $P = 4$. The noise power spectral density is equal to $N_0 = -174$ dBm/Hz and $P_{\mathrm{BS}} = 46$ dBm. The set $\mathcal{K}_b^{\mathrm{eMBB}}$ is determined using a proportional-fairness scheduler and transmission to co-scheduled UTs is done using a zero-forcing (ZF) MIMO precoder that can support up to $N = 4$ spatial layers.

We set $\mathrm{SINR}^{\mathrm{tr}} = 0$ dB if $N_p^{\mathrm{mini}} > 0$ and $\mathrm{SINR}^{\mathrm{tr}} = -2$ dB otherwise. This choice of values is motivated by the numerical example given at the end of Subsection III-D. Scheduling of URLLC packets is done using three methods, namely the proposed "precoder compatibility based preemptive scheduling", the ideal "precoder updating based preemptive scheduling" and "null-space-based preemptive scheduling" (NSBPS) from [3]. The second method is ideal in the sense that, contrary to the proposed scheme, MIMO precoders for the scheduled URLLC UT and the non-preempted eMBB UTs are computed from scratch during the mini-slot occupied by the URLLC transmission (rather than being kept unchanged) while at the same time assuming that such computation does not entail any additional processing delay. As for NSBPS, it is a state-of-the-art variant of preemptive scheduling in which all ongoing eMBB transmissions

[1] In some special cases, a device-specific parameter could additionally be used to generate the symbols. However, this parameter takes one of only few possible values, thus making it possible for the scheme to continue to function simply by means of trying out all these (few) possible values.

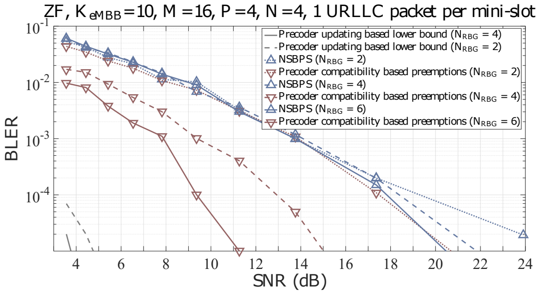

Fig. 2. BLER performance for URLLC traffic

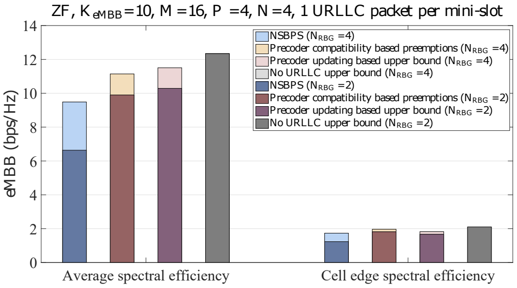

Fig. 3. Spectral efficiency performance for eMBB traffic

on the selected RBG are partially preempted as follows: upon the arrival of a URLLC packet, the scheduler picks the RBG on which the MIMO precoders of the eMBB transmissions are the most close to a predefined reference spatial subspace. Then, these precoders are projected on-the-go during the URLLC occupied mini-slot onto the reference subspace, while the precoder vector and the decoder matrix of the paired URLLC user are oriented into a possible null space of the reference subspace. Values of $\mathrm{BLER}_k^{\mathrm{URLLC}}$ resulting from applying the above three schemes is compared in Figure 2 for three values of N_{RGB}, namely $N_{\mathrm{RGB}} = 2, 4$ and 6, and plotted as function of the nominal SNR defined for UT k as $\mathrm{SNR}_k \triangleq g_k P_{\mathrm{BS}} / \left(\sigma_k^2 N\right)$. As expected, having more candidate RBGs for URLLC transmission improves the BLER performance. Indeed, with increasing values of N_{RGB} the chances of finding ongoing eMBB transmissions with precoders aligned closely enough with the reference spatial subspace (for NSBPS) or with the URLLC UT channel vector (for the proposed method) increase as well. However, results show that "precoder updating based preemptive scheduling" benefits more from larger values of N_{RGB}.

In Figure 3, eMBB rate performance of the proposed scheme is compared to that of NSBPS and to two upper bounds, namely the performance achieved by "precoder updating based preemptive scheduling" and by PF scheduling in the absence of URLLC traffic. This comparison is done using two metrics, namely the average spectral efficiency R_t^{eMBB} and cell edge spectral efficiency, i.e., R_t^{eMBB} computed when eMBB UTs pathloss coefficients take the value associated with the cell edge. It is clear that precoder compatibility based preempting causes smaller eMBB spectral efficiency loss than the loss due to NSBPS. This is because in NSBPS, *all* eMBB transmissions on the selected RBG necessarily suffer from projecting their respective MIMO precoders onto an arbitrary reference spatial direction, while in our scheme it is by design possible that only one ongoing eMBB transmission is paused.

V. Conclusions

We have proposed a new preemptive scheduler that minimizes both the MU-MIMO precoder computation time during mini-slots with URLLC packets and the downlink pilot overhead needed to coherently decode those packets. The key idea is that upon arrival, a URLLC user can replace the ongoing eMBB transmission whose MIMO precoder is the most compatible with the URLLC channel, i.e., whose channel vector is the most similar to the URLLC one. This allows avoiding recomputing the MU-MIMO precoding matrix while promptly serving the URLLC user. finally, the precoder compatibility metric upon which the scheduler is based can be acquired at the user terminal side in an opportunistic manner with minimal pilot overhead.

References

[1] A. Destounis, G. Paschos, J. Arnau, and M. Kountouris, "Scheduling URLLC users with reliable latency guarantees," in *Proc. WiOpt'18*, Shanghai, China, May 2018.

[2] H. Ji, S. Park, J. Yeo, Y. Kim, J. Lee, and B. Shim, "Ultra-Reliable and Low-Latency Communications in 5G Downlink: Physical Layer Aspects," *IEEE Commun. Mag.*, vol. 25, no. 3, pp. 124-130, June 2018.

[3] 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 2018*, Abu Dhabi, UAE, Dec. 2018.

[4] 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, June 2018.

[5] A. Anand, G. De Veciana, and S. Shakkottai, "Joint Scheduling of URLLC and eMBB Traffic in 5G Wireless Networks," in *Proc. IEEE INFOCOM 2018*, Honolulu, HI, USA, Apr. 2018.

[6] I. Parvez, A. Rahmati, I. Guvenc, A. I. Sarwat, and H. Dai, "A Survey on Low Latency Towards 5G: RAN, Core Network and Caching Solutions," to appear in *IEEE Communications Surveys & Tutorials*, 2018.

[7] M. Merda, A. W. Eckford, and R. Adve, "Updating Beamformers to Respond to Changes in Users," *arXiv preprint: 1803.04038*, Mar. 2018.

[8] H. Q. Ngo, T. L. Marzetta, and E. G. Larsson, "Analysis of the Pilot Contamination in Very Large Multicell Multiuser MIMO Systems for Physical Channel Models," in *Proc. IEEE ICASSP*, Prague, May 2011.

[9] F. Khan, *LTE for 4G Mobile Broadband: Air Interface Technologies and Performance*. Cambridge University Press, 2009.

[10] V. S. Abhayawardhana, I. J. Wassell, D. Crosby, M. P. Sellars, and M. G. Brown, "Comparison of Empirical Propagation Path Loss Models for Fixed Wireless Access Systems," in *Proc. IEEE VTC Spring*, Stockholm, Sweden, May 2005.