Preemption-Aware Rank Offloading Scheduling For Latency Critical Communications in 5G Networks

Abdul-Mawgood Ali Ali Esswie, Ali; Pedersen, Klaus Ingemann; Mogensen, Preben Elgaard

Published in:
2019 IEEE 89th Vehicular Technology Conference: VTC2019-Spring

DOI (link to publication from Publisher):
10.1109/VTCSpring.2019.8746364

Publication date:
2019

Document Version
Accepted author manuscript, peer reviewed version

Preemption-Aware Rank Offloading Scheduling For Latency Critical Communications in 5G Networks

Ali A. Esswie[1,2], Klaus I. Pedersen[1,2], and Preben E. Mogensen[1,2]
[1]Nokia Bell-Labs, Aalborg, Denmark
[2]Department of Electronic Systems, Aalborg University, Denmark

***Abstract*—This paper introduces a preemptive rank offloading scheduling framework for joint ultra-reliable low-latency communications (URLLC) and enhanced mobile broadband (eMBB) traffic in 5G new radio (NR). Proposed scheduler dynamically adapts the overall system optimization among the network-centric ergodic capacity and the user-centric URLLC one-way latency, based on the instantaneous traffic and radio resources availability. The spatial degrees of freedom, offered by the transmit antenna array, are fully exploited to maximize the overall spectral efficiency. However, when URLLC traffic buffering is foreseen, proposed scheduler immediately enforces scheduling pending URLLC payloads through preemption-aware subspace projection. Compared to the state-of-the-art schedulers from industry and academia, proposed scheduler framework shows significant scheduling flexibility in terms of the overall ergodic capacity and URLLC latency performance. The presented results therefore offer valuable insights of how to most efficiently multiplex joint URLLC-eMBB traffic over the 5G NR spectrum.**

***Index Terms*— URLLC; eMBB; 5G; MU-MIMO; New radio; Preemptive; Scheduling.**

I. Introduction

The coexistence of conventional human-centric and future machine-centric communications introduces more complex wireless environments [1, 2]. To address such diversified requirements, the standardization of the fifth generation new radio (5G-NR) is readily advancing, with its first specifications issued recently [3, 4]. 5G-NR features two major service classes: ultra-reliable low-latency communications (URLLC) and enhanced mobile broadband (eMBB). URLLC services require stringent latency and reliability targets, i.e., up to one-way radio latency of $1\,\text{ms}$ with 10^{-5} outage probability while eMBB applications seek for broadband data rates [5].

The efficient multiplexing of such diverse quality of service (QoS) classes over a single radio spectrum is a challenging and non-trivial scheduling problem, due to the underlying trade-off between latency, reliability, and aggregated data rate [6]. That is, if the system is forcibly engineered to satisfy the URLLC per-user outage of interest, the eMBB spectral efficiency (SE) will be severely degraded due to the inefficient resource utilization.

Recently, the URLLC and eMBB multiplexing problem has gained growing research attention from academia and industry. Primarily, the variable transmission time interval (TTI) duration with small data payloads is of significant importance to achieve the URLLC targets; however, at the expense of additional signaling overhead [7]. Spatial diversity techniques and dual connectivity [5] are also proved beneficial to improve the URLLC decoding ability by preserving the minimum outage signal-to-interference-noise-ratio (SINR). Furthermore, puncturing scheduler (PS) [8] is a state-of-the-art scheduling technique for joint URLLC-eMBB traffic, where the URLLC scheduling queuing delay becomes independent from the eMBB offered load through disruptive URLLC transmissions over eMBB-monopolized resources.

In our recent study [9], we demonstrated that a standard multi-user multi-input multi-output (MU-MIMO) transmission between URLLC-eMBB pairs is a fair solution to trade-off URLLC latency with overall SE. However, when the system spatial degrees of freedom (SDoFs) are limited, significant URLLC queuing delays are observed since a standard MU-MIMO pairing is only constrained by the achievable sum rate. Hence, in [10], we proposed a biased, and non-transparent version of the standard URLLC-eMBB MU-MIMO to guarantee an immediate and interference-free URLLC scheduling, regardless of the instantaneous system SDoFs and user loading. Thus, the URLLC latency budget is always preserved.

Compared to recent URLLC scheduler proposals, the scheduler operation is monotonically dictated by the URLLC capacity of interest. Examples include URLLC resource pre-allocation, and immediate puncturing. Thus, when URLLC services are multiplexed with eMBB applications on the same spectrum, the maximum system SE becomes infeasible. Needless to say, a multi-QoS-aware scheduling framework, which flexibly adapts the scheduling objectives to the instantaneous traffic state and being able to instantly preempt a particular QoS enforcement, is vital for future 5G-NR use cases.

In this work, we propose a preemption-aware rank offloading scheduling (PAROS) for joint URLLC and eMBB traffic. The proposed scheduler is a multi-objective framework, where both eMBB and URLLC QoS classes are simultaneously optimized on the TTI-level. Proposed PAROS scheduler first targets achieving the maximum possible ergodic capacity by attempting greedy MU eMBB transmissions. However, in case URLLC buffering is foreseen, hence, exceeding the critical URLLC latency budget, the PAROS scheduler enforces an instant subspace-projection for an interference-free URLLC scheduling over shared resources with ongoing eMBB transmissions. If the instantly available SDoFs are limited, the PAROS scheduler enforces an instant SDoF-relaxation through rank offloading, sufficient enough to immediately accommodate the incoming URLLC traffic. Hence, proposed scheduler shows great multiplexing flexibility in terms of the overall ergodic capacity and URLLC latency & reliability targets.

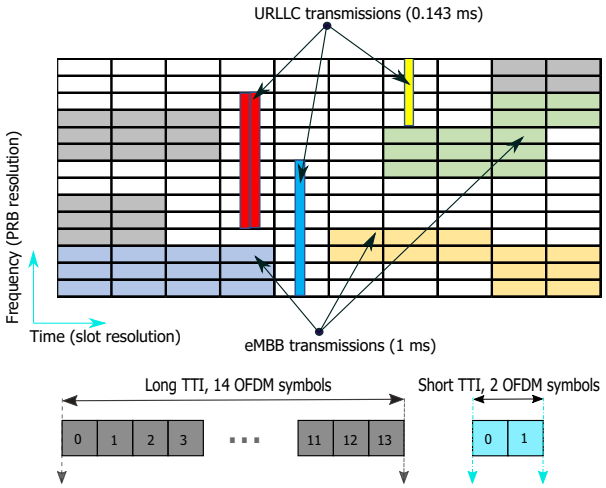

Fig. 1. Agile 5G-NR frame design and resource allocation.

Due to the complexity of the the 5G-NR scheduling problem [3] and addressed issues herein, we assess the performance of the proposed solution using extensive system level simulations, where the major scheduling functionalities are calibrated against the 3GPP 5G-NR assumptions. This includes the 3D channel spatial modeling, dynamic link adaptation, hybrid automatic repeat request (HARQ), dynamic multi-traffic modeling, SINR combining, and dynamic user scheduling.

This paper is organized as follows. System model is presented in Section II. The proposed scheduler framework is introduced in Section III while Section IV shows the numerical results. Finally, the paper is concluded in Section V.

II. System Model

We adopt a 5G-NR system with C downlink (DL) base-stations (BSs), each equipped with N_t transmit antennas. Each BS serves an average K uniformly distributed user equipment's (UEs), each with M_r receive antennas and $K = K_{\text{llc}} + K_{\text{mbb}}$, with K_{llc} and K_{mbb} as the average numbers of the URLLC and eMBB UEs per cell. Thus, the average cell loading condition per BS is defined by $\Omega = (K_{\text{mbb}}, K_{\text{llc}})$. The URLLC traffic is characterized by the FTP3 traffic model with a finite B-byte payload size and a Poisson point arrival process λ, while eMBB traffic is full buffer with infinite payload, to offer all-time best effort background load.

The 5G-NR flexible frame design is assumed. As depicted in Fig. 1, in the time domain, the URLLC traffic is scheduled over short TTI durations of 2-OFDM symbol mini slots, to satisfy its stringent latency budget. The eMBB traffic is scheduled over longer TTI durations of 14-OFDM symbol slots, to maximize the overall ergodic capacity. Furthermore, in line with [5], the scheduling grant is appended prior to the radio resources of the data payloads, thus, the minimum resource allocation per UE should be sufficiently large to accommodate both data and control symbols. In the frequency domain, the UEs are dynamically multiplexed by the orthogonal frequency division multiple access, where the smallest scheduling unit is the physical resource block (PRB) of 12-subcarriers.

We further assume a throughput-greedy scheduler with controlled, biased and non-transparent MU-MIMO transmissions, where a subset of co-scheduled UEs $\mathtt{G}_c \subseteq \mathcal{K}_c$ is allowed over an arbitrary PRB, where $\mathcal{K}_c$ is the active UE set in the c^{th} cell, $G_c = \mathbf{card}(\mathtt{G}_c)$, $G_c \leq N_t$ is the actual number of co-scheduled UEs and $\mathbf{card}(\cdot)$ indicates the cardinality. The post-decoded DL signal at the k^{th} UE from the c^{th} cell is given by

$$\hat{s}^{\kappa}_{k,c} = \left(\mathbf{u}^{\kappa}_{k,c}\right)^{\text{H}} \mathbf{H}_{k,c}\mathbf{v}^{\kappa}_{k,c}s_{k,c} + \sum_{j=1,j\neq c}^{C}\sum_{g\in\mathtt{G}_j}\left(\mathbf{u}^{\kappa}_{k,c}\right)^{\text{H}}\mathbf{H}_{k,j}\mathbf{v}_{g,j}s_{g,j}$$

$$+\begin{cases}\sum\limits_{g\in\mathtt{G}_c,g\neq k}\left(\mathbf{u}^{\kappa}_{k,c}\right)^{\text{H}}\mathbf{H}_{k,c}\mathbf{v}^{\{\text{'llc','mbb'}\}}_{g,c}s_{g,c}, & \kappa=\{\text{'mbb'}\}\\ \sim 0, & \kappa=\{\text{'llc'}\}\end{cases} + \mathbf{n}^{\kappa}_{k,c}, \tag{1}$$

where $\mathcal{X}^{\kappa}, \kappa\in\{\text{'llc', 'mbb'}\}$ denotes the QoS type requested by UE $\mathcal{X}$, $\mathbf{H}_{k,c} \in \mathcal{C}^{M_r\times N_t}, \forall k \in \{1,\ldots,K\}, \forall c \in \{1,\ldots,C\}$ follows the 3GPP 3D spatial channel [11] from the c^{th} cell to the k^{th} UE, $\mathbf{v}_{k,c} \in \mathcal{C}^{N_t\times 1}$ is the standard zero-forcing precoding vector, assuming a single stream transmission, and is expressed by

$$\mathbf{v}_{k,c} = \left(\mathbf{H}_{k,c}\right)^{\text{H}}\left(\mathbf{H}_{k,c}\left(\mathbf{H}_{k,c}\right)^{\text{H}}\right)^{-1}. \tag{2}$$

$s^{\kappa}_{k,c}$, $\hat{s}^{\kappa}_{k,c}$, and $\mathbf{n}^{\kappa}_{k,c} \in \mathcal{C}^{M_r\times 1}$ are the transmitted symbol, decoded symbol and the additive white Gaussian noise, respectively, while $\mathbf{u}^{\kappa}_{k,c}$ is the corresponding linear minimum mean square error interference rejection and combining (LMMSE-IRC) receiver matrix [5], with $(\cdot)^{\text{H}}$ as the Hermitian operation. The first summation in eq. (1) models the inter-cell inter-user interference, resulting from either URLLC or eMBB traffic while the second summation represents the intra-cell inter-user interference resulting from the overloaded MU-MIMO transmissions. As will be discussed in Section III, the URLLC-eMBB MU pairing is biased and altered such that URLLC traffic experiences no inter-user interference, hence, fulfilling its latency and reliability limits.

III. Proposed PAROS Scheduler

A. Problem Formulation

Multiplexing of the URLLC and eMBB QoS classes over the same radio spectrum implies a hard scheduling problem. URLLC QoS class must satisfy its outage of interest while eMBB QoS shall align with the network-wide outage. In that sequel, there is a trade-off between the user-centric URLLC and the network-centric eMBB targets. These are highly coupled and must be simultaneously optimized, i.e., eMBB rate maximization, and URLLC latency minimization as

$$\forall k_{\text{mbb}} \in \mathcal{K}_{\text{mbb}} : R_{\text{mbb}} = \underset{k_{\text{mbb}}\in\mathcal{K}_{\text{mbb}}}{\arg\max}\sum_{k_{\text{mbb}}=1}^{K_{\text{mbb}}}\sum_{rb\in\Xi^{\text{mbb}}_{k_{\text{mbb}}}}\beta_{k_{\text{mbb}}}r^{\text{mbb}}_{k_{\text{mbb}},rb}, \tag{3}$$

$$\forall k_{\text{llc}} \in \mathcal{K}_{\text{llc}} : \underset{k_{\text{llc}}\in\mathcal{K}_{\text{llc}}}{\arg\min}\left(\Psi_{k_{\text{llc}}}\right), \tag{4}$$

where $\forall k_{\text{mbb}} \in \{1,\ldots,K_{\text{mbb}}\}, \forall k_{\text{llc}} \in \{1,\ldots,K_{\text{llc}}\}$, R_{mbb} is the overall eMBB ergodic capacity, $\mathcal{K}_{\text{mbb}}$ and $\mathcal{K}_{\text{llc}}$ are the active UE sets of eMBB and URLLC QoS classes, respectively, $\Xi^{\text{mbb}}_{k_{\text{mbb}}}$ and $\beta_{k_{\text{mbb}}}$ imply the allocated set of PRBs and

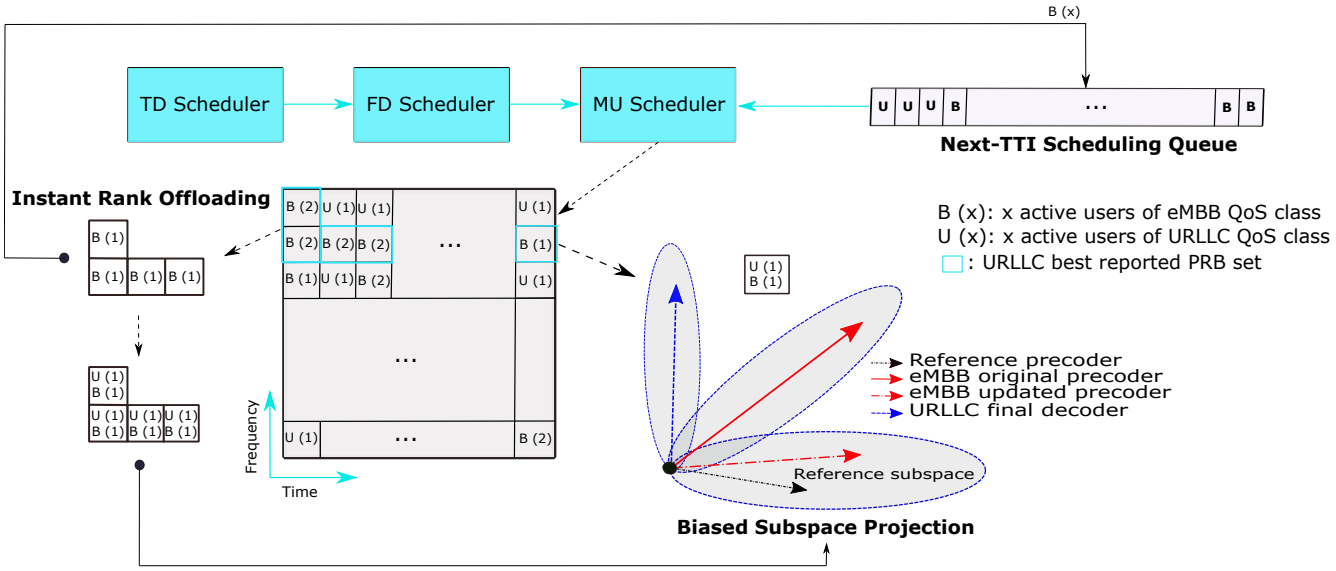

Fig. 2. Illustration example of the proposed PAROS scheduling framework with $G_c = 2$.

the scheduling priority of the k^{th} eMBB user. $r^{\text{mbb}}_{k_{\text{mbb}},rb}$ is the achievable k^{th} eMBB UE rate per PRB and $\Psi_{k_{\text{llc}}}$ is defined as the URLLC radio latency, as

$$\Psi_{k_{\text{llc}}} = \Lambda_{\text{q}} + \Lambda_{\text{bsp}} + \Lambda_{\text{fa}} + \Lambda_{\text{tx}} + \Lambda_{\text{uep}} + \Lambda_{\text{harq}}, \tag{5}$$

where $\Lambda_{\text{q}}, \Lambda_{\text{bsp}}, \Lambda_{\text{fa}}, \Lambda_{\text{tx}}, \Lambda_{\text{uep}}$ and Λ_{harq} are random variables to represent the queuing, BS processing, frame alignment, transmission, UE processing, and HARQ re-transmission delays, respectively. Λ_{fa} is upper bounded by the short TTI duration due to the agile 5G-NR frame structure, while the standardization bodies agreed that Λ_{bsp} and Λ_{uep} are each bounded by 3-OFDM symbol duration [5], because of the enhanced processing capabilities that come with the 5G-NR. Therefore, Λ_{tx}, Λ_{q} and Λ_{harq} are the major delay sources against achieving the URLLC latency deadline.

Therefore, to guarantee the URLLC radio latency limit, the URLLC traffic must fulfill: 1) not being buffered/queued over many TTI instances at the BS scheduler, and 2) one-shot transmissions without segmentation, to further allow for additional Λ_{harq} delay within the 1 ms deadline. This can be achieved by allocating excessive bandwidth for URLLC traffic, and enforcing a hard-coded URLLC higher priority in the scheduling buffers. As a result, the eMBB utility in (3) will be severely under-optimized, leading to a significant degradation of the overall SE. In that sequel, we address such multiplexing problem by proposing an efficient and flexibly adaptive scheduling framework.

B. Proposed Multi-Traffic PAROS Scheduler

The proposed scheduler dynamically alternates the scheduling targets in time such that the network ergodic capacity is maximized at all times by attempting greedy eMBB-eMBB MU-MIMO transmissions. When URLLC traffic buffering is foreseen, i.e., URLLC payload could not get scheduled from the time and frequency domain (TD, FD) schedulers, the proposed scheduler utilizes all system available SDoFs to instantly schedule these URLLC payloads over shared resources with transmitting eMBB UE through interference-free subspace projection based pairing. If the system PRBs are overloaded by eMBB MU transmissions, i.e., the maximum allowed number of per-PRB active users G_c is reached, PAROS scheduler immediately enforces eMBB UE offloading to reach $G_c - 1$ active UEs on the best reported PRBs of these incoming URLLC UEs. Fig. 2 shows an example of the proposed PAROS scheduler with $G_c = 2$.

At the BS – Time and frequency domain schedulers:

During an arbitrary TTI, if there is no sporadic URLLC traffic, PAROS framework allocates single-user (SU), i.e., rank-1, dedicated resources to newly arrived and/or buffered eMBB traffic, based on the standard proportional fair (PF) criterion over both TD and FD schedulers as

$$\Theta\left\{\text{PF}_{k_{\text{mbb}}}\right\} = \frac{r^{\text{mbb}}_{k_{\text{mbb}},rb}}{\overline{r}^{\text{mbb}}_{k_{\text{mbb}},rb}}, \tag{6}$$

$$k^{*}_{\text{mbb}} = \underset{k_{\text{mbb}} \in \mathcal{K}_{\text{mbb}}}{\arg\max} \ \Theta\left\{\text{PF}_{k_{\text{mbb}}}\right\}, \tag{7}$$

where $\overline{r}^{\text{mbb}}_{k_{\text{mbb}},rb}$ is the average received rate of the k^{th}_{mbb} UE. If URLLC payloads are available in the TD scheduling buffers, PAROS scheduler instantly overpowers the eMBB TD scheduling priority by the weighted PF criterion 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 instant URLLC scheduling. Then, the non-biased PF criterion is still applied on the FD scheduler to preserve fairness across the radio PRBs.

At the BS – Multi-user scheduler:

The PAROS scheduler aims to maximize the overall SE by default. Thus, at the MU scheduler, it always attempts greedy eMBB-to-eMBB MU transmissions, where G_c eMBB UEs are co-scheduled on an active PRB if the achievable sum rate is larger than that is of the primary eMBB UE only. In that sequel, the system PRBs are fully utilized with eMBB MU transmissions.

However, under high offered cell load, the schedulable resources may not be instantly available for critical URLLC

traffic. Thus, TD and FD schedulers fail to immediately schedule such traffic and it will be queued in the MU scheduling buffers. Then, PAROS first attempts a highly conservative MU transmission between a primary eMBB and secondary URLLC UE pair if their corresponding transmissions satisfy:

$$1 - \left| \left(\mathbf{v}^{\text{mbb}}_{k_{\text{mbb}}} \right)^{\text{H}} \mathbf{v}^{\text{llc}}_{k_{\text{llc}}} \right|^2 \geq \gamma. \tag{8}$$

The highly conservative, i.e., large, orthogonality threshold γ is enforced to protect the URLLC traffic against potential inter-user interference from the co-scheduled eMBB UE. If such orthogonality can not be offered at the current TTI, due to limited SDoFs, URLLC traffic shall be queued. Under this scheduling state, PAROS instantly alters the system optimization towards the URLLC latency and reliability targets instead of the ergodic capacity by satisfying the following conditions:

$$\textbf{rank} \left\{ \left(\mathbf{u}^{\text{llc}}_{k_{\text{llc}}} \right)^{\text{H}} \mathbf{H}_{k_{\text{llc}}} \mathbf{v}^{\text{llc}}_{k_{\text{llc}}} \right\} \sim \text{full}. \tag{9}$$

$$\textbf{rank} \left\{ \left(\mathbf{u}^{\text{llc}}_{k_{\text{llc}}} \right)^{\text{H}} \mathbf{H}_{k_{\text{llc}}} \mathbf{v}^{\text{mbb}}_{k_{\text{mbb}}} \right\} \sim 0. \tag{10}$$

Hence, PAROS scheduler instantly applies a biased and user-centric URLLC-eMBB MU transmission for interference-free URLLC scheduling, through subspace projection over the best reported URLLC PRBs with less than G_c active UEs. If such requested PRBs are overloaded with G_c eMBB active UEs, PAROS instantly offloads the eMBB UEs with the lowest achievable rates to preemptively free some SDoFs for URLLC traffic, i.e., it offloads PRBs with MU rank = G_c eMBB UEs down to $G_c - 1$ and biasedly pairs the incoming URLLC UE over these PRBs. Suspended eMBB transmissions are placed in the scheduling buffers according to their respective PF metrics. Furthermore, BS signals these eMBB UEs with a single-bit transmission interruption indication, for them to be aware that prior DL grant is not currently valid.

Towards such biased URLLC-eMBB pairing over an arbitrary PRB, a spatial reference subspace is predefined using the beamformed discrete Fourier transform, pointing to an arbitrary spatial direction θ, given by

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

where $(\cdot)^{\text{T}}$ implies the transpose operation and Δ is the antenna inter-distance. Then, PAROS scheduler searches for the active PRBs, from within the best reported PRB set of the incoming URLLC UEs, with at maximum $G_c - 1$ eMBB active UEs and whose active transmissions are closest possible in the spatial domain to the reference subspace as

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

where the Chordal distance between $\mathbf{v}^{\text{mbb}}_{k_{\text{mbb}}}$ and $\mathbf{v}_{\text{ref}}$ is given by

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

Finally, PAROS spatially projects the transmission of each victim eMBB UE $\mathbf{v}^{\text{mbb}}_{k^{\diamond}_{\text{mbb}}}$ over selected PRBs onto $\mathbf{v}_{\text{ref}}$ as

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

where $\mathcal{X} \cdot \mathcal{Y}$ indicates the dot product of $\mathcal{X}$ and $\mathcal{Y}$ and $\left(\mathbf{v}^{\text{mbb}}_{k^{\diamond}} \right)'$ is the post-projection precoder of the victim eMBB UE. Next, PAROS forcibly pairs incoming URLLC UEs over these shared resources with selected eMBB UEs. As the impacted eMBB UEs are not aware of the instant projection, eMBB capacity shall be degraded. However, due to the constraints in (8) and (12), the eMBB capacity is limited specially under high offered eMBB load, i.e., PAROS scheduler has a higher probability to fetch an eMBB UE whose transmission is originally aligned with the reference subspace, hence, the hard-coded spatial projection would not significantly degrade its achievable capacity. Furthermore, in our recent study [5], we have analytically determined that for a generic eMBB transmission, the loss function of the effective channel gain due to such spatial projection is scaled down by $\sin(\Phi)^2 \ll 1$, where Φ is the difference angle between pre-projection $\mathbf{v}^{\text{mbb}}_{k^{\diamond}_{\text{mbb}}}$ and post-projection $\left(\mathbf{v}^{\text{mbb}}_{k^{\diamond}_{\text{mbb}}} \right)'$ transmissions, leading to a guaranteed minimum loss rate. The BS scheduler finally signals the intended URLLC UEs with a single-bit true indication $\alpha = 1$.

At the URLLC UE:

When a URLLC UE acknowledges $\alpha = 1$, it realizes that its DL grant is shared with an active eMBB UE and the corresponding interfering transmission is aligned within the reference subspace. Thus, it first designs its first-stage LMMSE-IRC standard decoding matrix as expressed by

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

where $(\cdot)^{-1}$ stands for the inverse operation, and the interference covariance matrix $\mathbf{W}$ is given as

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

where $\mathbb{E}\{\cdot\}$ is the statistical expectation, σ^2 is the estimation error variance, and I_{M_r} denotes an identity matrix of size $M_r \times M_r$. Then, the URLLC UE intentionally transfers the statistics of $\left(\mathbf{u}^{\text{llc}}_{k_{\text{llc}}} \right)^{(1)}$ to a possible null space of the inter-user interference effective channel $\mathbf{H}_{k_{\text{llc}}} \mathbf{v}_{\text{ref}}$ as

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

Hence, the final URLLC decoding matrix $\left(\mathbf{u}^{\text{llc}}_{k_{\text{llc}}} \right)^{(2)}$ shall experience an interference-free transmission, leading to an improved URLLC decoding ability.

C. Comparsion to the state of the art URLLC schedulers

In this sub-section, we introduce the state-of-the-art scheduling proposals from both industry and academia, to which we compare the performance of proposed PAROS against.

Null space based preemptive scheduler (NSBPS) [10]: in our previous contribution, we proposed a monotonic scheduling optimization such that when URLLC queuing is inevitable,

Table I
MAJOR SIMULATION PARAMETERS.

Parameter	Value
Environment	3GPP-UMA,7 BSs, 21 cells
Channel bandwidth	10 MHz, FDD
Antenna setup	BS: 8 Tx, UE: 2 Rx
User load	$K_{\text{llc}} = 5$ or $20, K_{\text{mbb}} = 5$ or 20
User receiver	LMMSE-IRC
TTI configuration	URLLC: 0.143 ms (2 OFDM symbols) eMBB: 1 ms (14 OFDM symbols)
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: FTP3, B = 50 bytes, $\lambda = 250$ eMBB: full buffer
Multi-user rank	G_c = 2

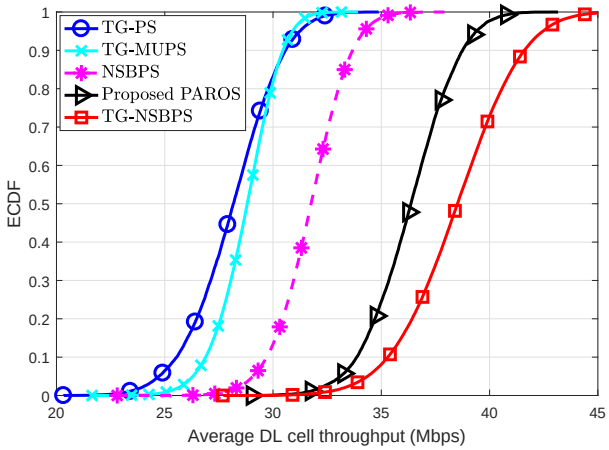

Fig. 3. Average cell throughput performance (Mbps).

the MU scheduler enforces a special URLLC-eMBB MU transmission, biased for the sake of the URLLC UEs. Hence, URLLC buffering is further minimized. However, eMBB-eMBB MU transmissions are not allowed to preserve the maximum possible SDoFs for incoming URLLC traffic.

Throughput-greedy NSBPS (TG-NSBPS): an extension of the NSBPS scheduler such that the scheduler always aims to maximizing the overall SE by attempting greedy eMBB-eMBB MU transmissions. When URLLC traffic is about to be buffered, TG-NSBPS instantly applies the NSBPS scheduling for immediate URLLC-eMBB MU pairing, however, only over the URLLC PRB set with less than G_c active eMBB UEs.

Throughput-greedy puncturing scheduler (TG-PS): an extension of the PS scheduler [8] where the MU scheduler always attempts greedy eMBB-eMBB MU transmissions in case there is no buffered URLLC traffic foreseen. Otherwise, to-be-buffered URLLC traffic preemptively overwrites some of the eMBB-monopolized PRBs for immediate scheduling, at the expense of the eMBB capacity degradation.

Throughput-greedy Multi-user PS (TG-MUPS): an extension to the MUPS scheduler in [9], in which the scheduler attempts greedy eMBB-eMBB MU transmissions if there is no URLLC queued traffic. In case URLLC traffic is to be buffered for multiple TTIs, scheduler attempts a *standard and non-biased* URLLC-eMBB MU transmissions based on the achievable sum rate constraint, only over the PRB set with maximum $G_c - 1$ eMBB active UEs. If a successful pairing is not possible, scheduler immediately rolls back to PS scheduler by overwriting several ongoing eMBB transmissions.

IV. NUMERICAL RESULTS

The performance evaluation is based on dynamic system level simulations where the 3GPP 5G-NR methdology is followed [5]. We adopt 8×2 antenna setup, with the 3D spatial channel modeling. Dynamic link adaptation and Chase combining HARQ are used to relax the initial block error rate (BLER). The main simulation settings are listed in Table I. Herein, we consider the NSBPS scheduler as a reference against other schedulers under evaluation.

Fig. 3 shows the empirical cumulative distribution function (ECDF) of the average DL cell throughput performance for all assessed schedulers with $\Omega = (5, 5)$. The NSBPS scheduler provides a fair cell throughput performance since all system SDoFs are fully reserved for instant URLLC scheduling, i.e., greedy eMBB-eMBB MU transmissions are not allowed regardless from the URLLC traffic availability. The proposed PAROS scheduler offers a significant improvement of the cell throughput, i.e., an average of 5 Mbps throughput increase compared to the NSBPS scheduler, while the TG-NSBPS scheduler offers the best cell throughput due to the aggressive MU transmissions without rank offloading.

Moreover, the TG-PS scheduler exhibits a severe degradation in the overall throughput due to the puncturing events. Thus, punctured eMBB transmissions suffer from significant capacity loss. Consequently, the SE gain from the greedy eMBB-eMBB MU pairings vanishes due to the puncturing capacity loss, e.g., one URLLC UE may puncture an active PRB with G_c active eMBB UEs, thus, degrading their respective capacity. Finally, the TG-MUPS shows a slightly improved ergodic capacity than the TG-PS due to the successful URLLC-eMBB MU standard pairings, hence, no puncturing is applied. Otherwise, TG-MUPS rolls back to PS scheduler for instant URLLC transmission.

As shown in Fig. 4, the empirical complementary CDF (ECCDF) of the URLLC radio latency is depicted. Referring to the NSBPS scheduler, the proposed PAROS, and TG-PS schedulers offer a decent URLLC latency performance, approaching its stringent target, i.e., 1 ms at 10^{-5} outage probability. Thus, if there is buffered URLLC traffic at the MU scheduler, which is the last scheduling opportunity for URLLC traffic to get scheduled during the current TTI, both schedulers enforce an *immediate and biased* URLLC transmissions regardless of the scheduler state. Thus, the URLLC queuing delay is significantly minimized. However, the TG-MUPS exhibits an increase of $\sim +43.4\%$ in the URLLC latency than the PAROS scheduler. This is basically due to the *standard and non-biased* URLLC-eMBB MU transmissions, where the resulting inter-user interference degrades the URLLC decoding ability, leading to several re-transmissions prior to a successful decoding. The TG-NSBPS shows the worst URLLC latency since all active PRBs are highly likely to be overloaded with

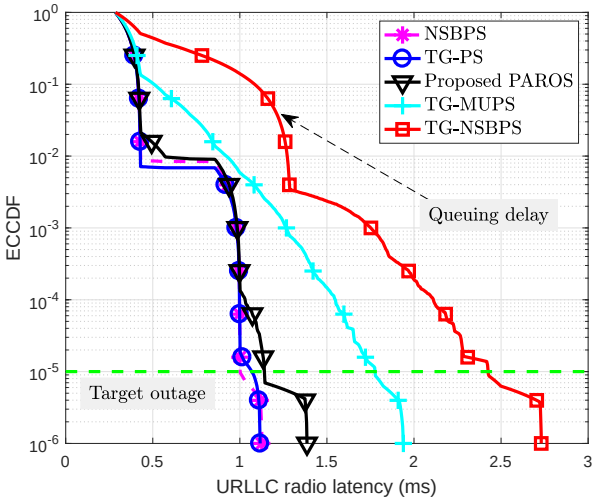

Fig. 4. URLLC latency performance (ms).

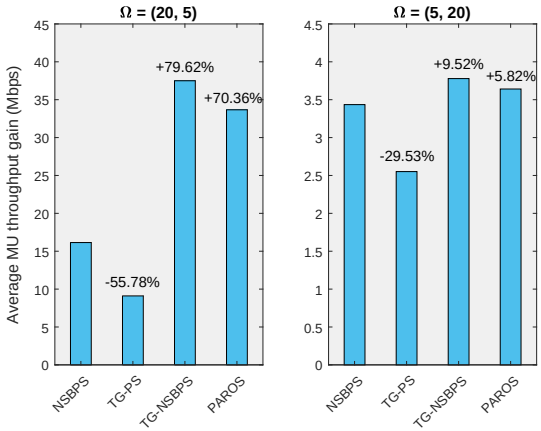

Fig. 5. MU throughput performance (Mbps), compared to NSBPS.

G_c active eMBB UEs. Thus, when URLLC traffic arrives the MU schedulers, it has very limited SDoFs to schedule such critical traffic, resulting in further URLLC queuing delays.

Finally, Fig. 5 presents a comparison of the achievable MU throughput increase, with respect to the SU case, for two extreme loading states. The MU achievable throughput is defined as the pre-detection sum data rate due to the effective MU pairings at the BS. Thus, for SDoF-rich state, i.e., $\Omega = (20, 5)$, where there is a sufficient number of active eMBB UEs, TG-NSBPS and PAROS schedulers offer a significant enhancement in the achievable MU throughput due to the successful eMBB-eMBB MU pairings. Thus, the ergodic capacity is almost doubled, i.e., $\geq +70\%$ gain. Though, PAROS scheduler exhibits $\sim -9.5\%$ MU loss than TG-NSBPS due to the instant rank offloading when URLLC buffering is envisioned. Finally, the TG-PS scheduler exhibits a severe degradation in the MU throughput since under such loading state, the majority of the system PRBs are overloaded with eMBB MU transmissions. Thus, instant puncturing of these becomes quite costly. With $\Omega = (5, 20)$, the system becomes dictated by URLLC transmissions from the TD and FD schedulers. Hence, all schedulers suffer from MU degradation since URLLC-URLLC MU transmissions are not allowed.

V. Concluding Remarks

We have proposed a preemption-aware rank offloading scheduling (PAROS) framework for 5G new radio. The proposed scheduler shows great scheduling flexibility in multi-traffic scenarios, i.e., URLLC and eMBB. It dynamically adapts the scheduling objectives according to the instantaneous traffic availability and scheduling state. Compared to the state-of-the-art scheduler proposals, the proposed PAROS scheduler offers a significantly improved ergodic capacity of more than 70% gain, while simultaneously satisfying the URLLC stringent latency and reliability targets, i.e., $1\,\text{ms}$ at 10^{-5} outage.

The valuable insights offered by this work are summarized as: (1) for highly loaded cells, multi-traffic spatial schedulers become of a significant importance to trade-off the overall spectral efficiency with the latency and reliability targets, (2) conventional spatial schedulers are not appropriate for latency critical URLLC traffic due to their network-centric, instead of user-centric, scheduling constraints, and (3) these schedulers should be sufficiently flexible to maximize the ergodic capacity by default and be able to preemptively free sufficient degrees of freedom for the sporadic URLLC arrivals. A further flexible URLLC-to-URLLC multi-user scheduling study will be conducted in a future work.

VI. Acknowledgments

This work is partly funded by the Innovation Fund Denmark, Grant: 7038-00009B. Also, part of this work is performed in the framework of the Horizon 2020 project ONE5G (ICT-760809) receiving funds from the European Union.

References

[1] R. Drath and A. Horch, "Industrie 4.0: hit or hype?," *IEEE Ind. Electron. Mag.*, vol. 8, no. 2, pp. 56-58, June 2014.

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

[3] Study on new radio access technology (Release 14), 3GPP, TR 38.801, V14.0.0, March 2017.

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

[5] A. A. Esswie and K. I. Pedersen, "Opportunistic spatial preemptive scheduling for URLLC and eMBB coexistence in multi-user 5G networks," *IEEE Netw.*, vol. 6, pp. 38451-38463, July 2018.

[6] B. Soret, P. Mogensen, K. I. Pedersen and M. C. Aguayo-Torres, "Fundamental trade-offs among reliability, latency and throughput," *in Proc. IEEE Globecom*, Austin, TX, 2014, pp. 1391-1396.

[7] K. Pedersen, G. Pocovi, J. Steiner and A. Maeder, "Agile 5G scheduler for improved E2E performance for different network implementations," *IEEE Commun. Mag.*, vol. 56, no. 3, pp. 210-217, Mar. 2018.

[8] 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.

[9] 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.

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

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