

Opportunistic Overlapping: Joint scheduling of uplink URLLC/eMBB traffic in NOMA based Wireless Systems

Arjun Anand*, Gustavo de Veciana[†], Derya Malak[‡], Ayman Elezabi[§], and Aniruddh Venkatakrishnan[†]

*Intel Corporation, USA, arjun.anand@intel.com

[†]ECE Dept., The University of Texas at Austin, USA, deveciana@utexas.edu, aniruddh.venkat@utexas.edu

[‡]Communication Systems Dept., EURECOM, 06904 Biot Sophia Antipolis cedex, FRANCE, derya.malak@eurecom.fr

[§]ECNG Dept., The American University in Cairo, Egypt, aelezabi@aucegypt.edu

Abstract—We consider the joint scheduling of uplink URLLC and eMBB user traffic in a cellular system. The central challenge is coordinating URLLC uplink user transmissions for traffic requiring extremely low latency, high reliability, and in the absence of knowledge of instantaneous URLLC channel qualities, albeit with knowledge of channel distributions. To avoid collisions and meet latency and reliability constraints, we propose to pre-optimize layouts of non-overlapping transmission opportunities for URLLC users, which may, or may not, be used depending on their traffic. To increase overall throughput we propose to leverage Non-Orthogonal Multiple Access (NOMA) based opportunistic scheduling of overlapping eMBB user traffic and propose power control policies, i.e. eMBB transmit power backoff, to protect possible URLLC transmissions from overlapping eMBB traffic. We derive the sum-rate optimal power control for eMBB traffic, and propose a linear approximation that simplifies the later scheduling task. Utilizing an outage capacity model for the unknown URLLC channel qualities, we assign the URLLC allocations as a greedy first fit decreasing packing problem. We then apply an opportunistic overlapping scheduler that, subject to meeting URLLC users' latency constraints, optimizes eMBB users' sum utility. Substantial discrete event simulations were conducted to explore the performance impact of system parameters associated with URLLC traffic requirements, eMBB power control, etc. Depending on the traffic scenarios, we show gains reaching 75% in the sum eMBB throughput and/or 5th percentile throughput relative to an orthogonal multiple access baseline.

Index Terms—wireless uplink scheduling, URLLC, eMBB, non-orthogonal multiple access

I. INTRODUCTION

The next generation of wireless standards, e.g., 5G and 6G, are being specifically designed to support a wider range of traffic types, e.g., Ultra Reliable Low Latency Communications (URLLC) to enhanced Mobile Broadband (eMBB) traffic, and of course meet particularly challenging performance guarantees. Multi-user scheduling of uplink URLLC users' transmissions is a particularly challenging task for two key reasons. First in order to achieve extremely low latency one must be able to schedule on timescales faster than those typically used in today's systems, i.e., the frame, so if a packet becomes available in the midst of a frame transmission it can be transmitted without delaying and scheduling it on the next

frame. A proposed solution to achieve this on the downlink is to allow puncturing or superposition of URLLC transmissions on previously scheduled eMBB traffic, with the possibility that eMBB users' transmissions can recover from puncturing or superposed transmissions through coding/HARQ and/or cancellation of URLLC 'interference' when decoding eMBB transmissions. This approach however can not be directly applied on the uplink, due to the second key challenge – since mobile devices are distributed, they need to coordinate to make sure they do not collide on uplink transmissions. While such coordination is straightforward on the downlink, on the uplink it would require timely exchanges amongst uplink users precluding the desired low latency.

One way to overcome such coordination delays is to simply pre-allocate uplink transmission *opportunities* for URLLC users. By allocating a sufficient number of, say periodic, transmission opportunities in frames to each user one can ensure they will have an opportunity to send a packet soon after it is available for transmission. Such an allocation can be viewed as a premium service level agreement. Of course if the user traffic is perfectly timed, the URLLC user can ensure it makes use of these pre-allocations. However in practice this is not the case, which may lead to inefficient resource usage, particularly since URLLC users require high reliability and hence will make use of a typically large number of subbands in a short period to achieve both high reliability and low transmission delay. Thus, it makes sense once again to consider overlapping preallocated URLLC transmission opportunities with scheduled eMBB traffic and make the most of non-orthogonal multiple access (NOMA) transmission modalities. This paper explores such a framework and studies the potential performance of joint scheduling and power control of uplink URLLC (transmission opportunities) and eMBB traffic in a NOMA based wireless system.

A. Contributions

The main contributions of this paper are in two areas:

1. *Framework and power control*: We develop a framework to study the joint layout and scheduling of uplink URLLC and eMBB user traffic leveraging NOMA-based techniques.

Critical to this approach is ensuring that URLLC transmissions are reliable and have low latency, and thus that a joint power control policy is developed to protect URLLC users from overlapping eMBB traffic as well as other cell interference. A power control policy in this setting corresponds to the choice of a tradeoff between overlapping URLLC and eMBB transmission rates. In Section 3 of this paper we propose and study a sum throughput (URLLC + eMBB) optimal power control strategy and approximations thereof that enable simpler approaches to overlapping resource allocations.

2. Joint scheduling: In order to study the resource savings of our NOMA based framework vs a purely OMA based approach, we propose an algorithm in Section 4 to layout persistent URLLC transmission opportunities across the frame in accordance to the derived power control strategy while opportunistically scheduling eMBB user traffic so as to optimize the sum utility of their allocated long term rates. Opportunism in this setting involves not only exploiting variations in channel qualities but also variations due to the presence of pre-allocated URLLC transmission opportunities. The latter is naturally coupled to the impact the overlap choices of users with different channel qualities will have on eMBB power control. *A key novelty of our work lies in exploring dynamic resource allocation/overlapping while maintaining latency and reliability objectives for heterogeneous URLLC and eMBB users, which have asymmetric knowledge of their current quality of their channels.* We develop both a well founded theoretical basis for our proposed algorithms and explore their performance via simulations showing scenario dependent eMBB sum throughput and 5th percentile user throughput gains over OMA up to 75%.

B. Related Work

There has been extensive related work yet to the best of our knowledge none addresses the joint dynamic optimization of URLLC placement and eMBB scheduling addressed in this paper.

Wireless Scheduling and network utility maximization. The design of opportunistic multi-user wireless schedulers for utility maximization, delay optimization and queue stability has received substantial attention in the last two decades. We refer to [1] and [2] and the references therein. Joint scheduling of different types of traffic flows in wireless networks using NOMA has been considered mainly for downlink, e.g., as in [3] by jointly deciding user selection, power allocation, Modulation and Coding Scheme (MCS) selection for NOMA-based downlink systems via centralized schedulers, or incorporating superposition/puncturing techniques, e.g., in [4], [5].

Existing work on joint scheduling primarily focuses on eMBB/URLLC downlink traffic. For example, in [6] the authors conducted dynamic resource optimization to meet heterogeneous service requirements. Researchers maximized the URLLC load without sacrificing the eMBB throughput, e.g., in [7] via simulation, and [8] by designing a resource allocation and a scheduling mechanism by puncturing the eMBB transmissions to guarantee a minimum achievable

eMBB rate. In [9] authors proposed a channel and QoS-aware dynamic resource allocation and joint link adaptation policy to multiplex traffic. By incorporating constraints on service isolation, latency, minimum rate, reliability, and adaptive modulation coding, in [10] authors optimized the sum-rate for a resource slicing setting. In [11] authors scheduled coexisting traffic in downlink 5G NR to enhance system throughput and provide URLLC delay guarantees. Preemptive scheduling for eMBB/URLLC downlink traffic was considered in [12], e.g., to offload eMBB and to improve ergodic capacity and or ensure URLLC latency as in [13], and for multi-user MIMO systems in [14]. In [15] and [16] authors considered a spatial preemptive scheduling for eMBB/URLLC in downlink 5G NR.

System-level simulations for uplink. In [17] the authors explore NOMA and adaptive power allocation and the realization of general fairness versus throughput tradeoffs for both downlink and uplink. Authors in [17] contrasted the resource requirement with fixed power [18] and cognitive radio-inspired [19] NOMA models, but did not optimize the aggregate resource requirement. There exist many approaches to multiplex eMBB/URLLC on the uplink, see e.g., [20]–[23], which aim to achieve high spectral efficiency by enabling grant-based and grant-free users to share the same spectrum resources and meet the strict latency and reliability constraints in 5G NR. Due to the complexity of the analysis, most of these evaluations, e.g., in [20]–[22], are based on system-level simulations.

II. SYSTEM MODEL

Uplink frame and traffic demands model. We shall consider an uplink frame for a wireless Base Station (BS) shared by a set of URLLC and eMBB users, denoted U and E respectively. The frame is assumed to have a width of w and height of h Resource Blocks (RBs) where each RB has a width of m minislots and height of n subcarriers. Thus the overall frame has a width of $f_w = w \times m$ minislots and height $f_h = h \times n$ subcarriers and its resources can be individually indexed as $F = \{1, 2, \dots, f_w\} \times \{1, 2, \dots, f_h\}$.¹

URLLC users require extremely low latency (lower than one frame period) and high reliability. Scheduling such uplink transmissions on a timescale shorter than a frame is difficult since resource allocations are made, at best, at the start of a frame and one needs to coordinate among uplink users to avoid collisions. Hence in order to meet such requirements for each user $u \in U$ one pre-allocates periodic *transmission opportunities* to send b_u bits every p_u minislots. The entire b_u bits should be sent within a given minislot. Thus if the the user's traffic were a periodic stream of b_u bits every p_u minislots the maximum delay would be at most p_u minislots. We refer to these as 'transmission opportunities' because they

¹In 3GPP standards, allocations are in terms of physical RBs (PRBs). Starting with Rel-15 [24], sub-PRBs were introduced, which may occupy 3, 6, or 9 subcarriers, and mini-slots, which may be 2, 4, or 7 OFDM symbols for sub-6 GHz bands. In this work, we consider fine granularity in resource allocations down to subcarriers and mini-slots for conceptual clarity, as well as the coarser granularity used in practice.

may or may not be used by the URLLC user. However, as we shall see, they may overlap with one or more scheduled eMBB user transmissions. In the sequel, Section IV, we shall assume eMBB user queues are infinitely backlogged and optimize their rate allocations for their sum utility.

Overlapping URLLC/eMBB user transmissions. We consider NOMA based sharing of uplink resources which permits overlapping of URLLC transmission opportunities and scheduled eMBB user transmissions, but no overlapping (collisions) of URLLC transmissions or overlapping among scheduled eMBB users transmissions. When URLLC/eMBB overlaps occur we propose to use Successive Interference Cancellation (SIC), i.e., we first decode URLLC users' data (treating overlapping eMBB transmissions as additive white Gaussian noise) and then if successful we subtract the decoded overlapping URLLC signal before proceeding to decode eMBB users' transmissions. Ensuring this process is successful with high probability requires a joint power control policy for overlapping transmissions, which factors their respective channel qualities, and that guarantees a high probability of correct detection for URLLC users' transmissions, thus avoiding further delays associated with HARQ. Finally to ensure high reliability for URLLC users and goodput for eMBB users' uplink transmissions, one must manage the uncertainty associated with inter-cell interference. Below, we develop a general framework to study the potential of enabling such overlapping.

URLLC and eMBB channel gains. The transmission rates of URLLC and eMBB users and overlaps thereof are driven by their respective channel gains. Multiple transmission opportunities for URLLC users are periodically allocated within a frame, during which the realizations of fast channel variations may not be known. Thus for URLLC user $u \in U$ we assume the distribution for the random channel quality is Q_u^U , which can be measured/modelled, accounting for possibly distance dependent path loss, shadowing and fast fading. We let $q_u^{U,\epsilon}$ be such that $P(Q_u^U > q_u^{U,\epsilon}) = 1 - \frac{\epsilon}{2}$, i.e., user u can count on a channel quality of at least $q_u^{U,\epsilon}$ with reliability $1 - \frac{\epsilon}{2}$. We refer to this as the user's 'reliable' channel quality. We thus apply an outage capacity model, thus avoiding the need to know instantaneous URLLC channel qualities. By contrast the channel quality of an eMBB user $e \in E$ denoted q_e^E is known at the start of the frame or earlier, as that information is used by the scheduler. Note, for simplicity, that we have assumed that both URLLC and eMBB users see flat fading across the subcarriers on a given frame, i.e., the channel gains are the same across the frame's RBs.

Background and inter-cell interference as noise. Uplink transmissions can be quite sensitive to inter-cell interference. In this paper we model it as a (conservatively selected) constant or more generally as a random variable O denoting the interference from other cells seen per subcarrier bandwidth at the BS under consideration. One can then define o^ϵ such that $P(O \leq o^\epsilon) = 1 - \frac{\epsilon}{2}$, i.e., with probability $1 - \frac{\epsilon}{2}$ one can be certain that the interference will not exceed o^ϵ , see e.g., [2]. We model the sum of inter-cell interference and thermal background noise at a BS as additive white Gaussian noise

giving a total power of $\sigma^2 = o^\epsilon + N_0W$ per subcarrier where W is the bandwidth per subcarrier and N_0 the noise density.

III. CAPACITY MODELS AND EMBB POWER CONTROL

We shall assume that if a URLLC user is active during its transmit opportunity it transmits at its maximum transmit power p_{\max} which it divides equally amongst the, say, k subcarriers it has been pre-allocated. The rationale for doing so is to maximize the likelihood the BS successfully decodes the URLLC transmissions, which can then be subtracted during eMBB decoding². In the sequel we shall let $\beta = \frac{p_{\max}}{\sigma^2}$ denote a normalized uplink signal power to background noise plus inter-cell interference power ratio.

Suppose an eMBB user with channel quality q_e^E overlaps with a transmission opportunity for a URLLC user channel quality $q_u^{U,\epsilon}$. To ensure the URLLC user is successfully decoded one must limit the overlapping eMBB user's power per subcarrier – we denote this upper bound by $p(q_u^{U,\epsilon}, q_e^E, \gamma)$. We will parameterize eMBB power control policies by selecting a function $\gamma(\cdot)$ depending, for example, on $q_u^{U,\epsilon}, q_e^E, p_{\max}, \sigma^2$ and which denotes the target SINR for URLLC user u when overlapping with eMBB user e . If the eMBB user e gets l subcarriers, the upper bound is then given by

$$p(q_u^{U,\epsilon}, q_e^E, \gamma) = \min \left[\max_{\rho \geq 0} \left[\rho \mid \frac{\frac{p_{\max}}{k} q_u^{U,\epsilon}}{\sigma^2 + \rho q_e^E} \geq \frac{\gamma}{k} \right], \frac{p_{\max}}{l} \right], \quad (1)$$

where we have suppressed the arguments of γ and assumed that a URLLC user that splits its power over k subbands, i.e., p_{\max}/k , and also scales his target SINR by γ/k – hence the k 's will cancel. The choice of γ embodies a tradeoff between URLLC and overlapping eMBB rates. At the end of this section we shall choose γ to maximize the sum rate of an URLLC/eMBB overlap and discuss approximations thereof, but for now we shall keep our framework generic.

Finally note that an eMBB user $e \in E$ with channel quality q_e^E may at a given time be allocated subcarriers which overlap with a set $U_e \subset U$ of URLLC users. In order to protect all URLLC users the eMBB user should use a power no larger than

$$\min_{u \in U_e} p(q_u^{U,\epsilon}, q_e^E, \gamma) = p(\min_{u \in U_e} q_u^{U,\epsilon}, q_e^E, \gamma),$$

where we have assumed $\gamma(\cdot)$ is such that $p(\cdot, q_e^E, \gamma)$ is non-decreasing in the first argument which requires that γ grows at most linearly in the same argument. Assuming an eMBB user allocates an equal share of its maximum power p_{\max} to its l subcarriers, the power per subcarrier is set to

$$\min \left[p(\min_{u \in U_e} q_u^{U,\epsilon}, q_e^E, \gamma), \frac{p_{\max}}{l} \right]. \quad (2)$$

Achieved capacity under NOMA/OMA. Consider a URLLC user with channel quality $q_u^{U,\epsilon}$ which is allocated k subcarriers and overlapping with one or more eMBB users. Our power

²In general this may be too aggressive, since it will impact neighboring BSs' uplink transmissions. Hence, one might wish to have a more friendly URLLC user uplink power control policy that depends on the users' channel quality.

control policy for eMBB users ensures the URLLC users see an SINR of at least $\frac{\gamma}{k}$ across subcarriers, and the maximum rate achievable by pre-allocating k subcarriers in a minislot for the URLLC user to transmit is

$$c_{u,e}^U(k) = k\mu \frac{W}{2} \log(1 + \frac{\gamma}{k}) \text{ bits/minislot}, \quad (3)$$

where μ is the minislot duration. Recall the function γ may depend on $q_u^{U,\epsilon}$ and q_e^E thus the capacity $c_{u,e}^U$ is indexed accordingly. Since this function is increasing in k , if user u requires an opportunity to transmit b bits on a minislot it will need to be allocated

$$k_{u,e}^U(b) = \lceil \min_k [k \mid c_{u,e}^U(k) \geq b] \rceil \quad (4)$$

subcarriers. The above capacity formulas are contingent on a sufficiently good URLLC channel and not seeing an excessively high intercell interference, for our system model this occurs with probability $(1 - \frac{\epsilon}{2})^2 \approx 1 - \epsilon$ which corresponds to the desired URLLC reliability. To keep things simple we have used the usual capacity formula leaving practical modulation and coding schemes and/or the impact of finite blocklengths aside, see e.g., [25].

If a URLLC user $u \in U$ with channel gain $q_u^{U,\epsilon}$ is assigned k subcarriers in a minislot without overlaps with eMBB users, i.e., orthogonal multiple access (OMA), we shall model and denote its capacity by

$$c_{u,0}^U(k) = k\mu \frac{W}{2} \log(1 + \frac{\beta}{k} q_u^{U,\epsilon}) \text{ bits/minislot}. \quad (5)$$

Once again this function is increasing in k thus as in the case with overlaps if user u has to transmit b bits on a minislot it will need to be allocated following no. of subcarriers:

$$k_{u,0}^U(b) := \lceil \min_k [k \mid c_{u,0}^U(k) \geq b] \rceil. \quad (6)$$

Now consider an eMBB user $e \in E$ with channel quality q_e^E . As mentioned earlier URLLC users will be successfully decoded and their interference will be cancelled with probability $(1 - \epsilon)$. Note that the URLLC users' channels are estimated for coherent detection but are not known prior to scheduling, as mentioned earlier. Thus, the average capacity (in bits/minislot) achieved by eMBB user e when it is allocated l subcarriers which overlap with the subset U_e of URLLC users is thus given by

$$c_{U_e,e}^E(l) = l(1 - \epsilon)\mu \frac{W}{2} \log(1 + \frac{\min[p(\min_{u \in U_e} q_u^{U,\epsilon}, q_e^E, \gamma), \frac{p_{\max}}{l}] q_e^E}{\sigma^2}) \quad (7)$$

If an eMBB user with channel quality q_e^E is assigned dedicated resources without overlaps with URLLC users, i.e., OMA, we shall model and denote its capacity as follows

$$c_{0,e}^E(l) = l\mu \frac{W}{2} \log(1 + \frac{\beta}{l} q_e^E) \text{ bits/minislot}. \quad (8)$$

Optimizing eMBB power control. As mentioned earlier γ , and thus eMBB power control, could depend on various characteristics of the overlapping transmissions and system

criteria - below we propose a sum-rate optimal power control policy and derive the γ that achieves it. We then propose a linear power control policy, which approximates the sum-rate optimal policy, and simplifies the subsequent scheduling stage.

Proposition 1. Sum-rate Optimal Power Control (OPC) policy. Consider an eMBB user with channel quality q_e^E which is allocated l subcarriers overlapping with a URLLC user with channel quality $q_u^{U,\epsilon}$ which has been allocated k subcarriers. We refer to the sum-rate optimal eMBB power control as setting the target γ in Eq. (1) so as to maximize the sum of the normalized URLLC/eMBB rates of the overlap. It can be shown to be independent of k and given by

$$\gamma^{OPC}(q_u^{U,\epsilon}, q_e^E, \beta, l) = \frac{\beta q_u^{U,\epsilon}}{1 + \frac{\beta}{l} q_e^E}. \quad (9)$$

This can be viewed as generalizing conventional channel inversion power control based on the eMBB user channel quality while including a linear scaling in URLLC users' channel quality.

Proof. From (3) and (7), the achievable overlap rate (normalized by $\mu \frac{W}{2}$) per subcarrier for the given pair of users is

$$\begin{aligned} & \frac{1}{\mu \frac{W}{2}} \left(\frac{c_{u,e}^U(k)}{k} + \frac{c_{U_e,e}^E(l)}{l} \right) = \log(1 + \frac{\gamma}{k}) \\ & + (1 - \epsilon) \log(1 + \frac{\min[p(\min_{u \in U_e} q_u^{U,\epsilon}, q_e^E, \gamma), \frac{p_{\max}}{l}] q_e^E}{\sigma^2}) \\ & \stackrel{(a)}{=} (1 - \epsilon) \log(1 + \frac{\gamma}{k})^{\frac{1}{1-\epsilon}} \\ & + (1 - \epsilon) \log(1 + \frac{1}{\sigma^2} \min[\frac{p_{\max} q_u^{U,\epsilon}}{\gamma} - \sigma^2, \frac{p_{\max} q_e^E}{l}]) \\ & \stackrel{(b)}{=} \log(1 + \frac{\gamma}{k}) + (1 - \epsilon) \log(\min[\frac{p_{\max} q_u^{U,\epsilon}}{\gamma \sigma^2}, 1 + \frac{p_{\max} q_e^E}{l \sigma^2}]) \\ & \stackrel{(c)}{=} (1 - \epsilon) \min \left[\log \left(\frac{(1 + \frac{\gamma}{k})^{\frac{1}{1-\epsilon}}}{\gamma} \beta q_u^{U,\epsilon} \right), \right. \\ & \quad \left. \log \left(\left(1 + \frac{\gamma}{k}\right)^{\frac{1}{1-\epsilon}} \left(1 + \frac{\beta}{l} q_e^E\right) \right) \right], \quad (10) \end{aligned}$$

where (a) follows from the fact that if (1) has a feasible solution, it yields $p(q_u^{U,\epsilon}, q_e^E, \gamma) q_e^E = \min[\frac{p_{\max} q_u^{U,\epsilon}}{\gamma} - \sigma^2, p_{\max} q_e^E]$. Simplifying the second term in (a) yields (b), and (c) follows from interchanging the order of the min() and log() due to monotonicity of the log() function and substituting $\beta = \frac{p_{\max}}{\sigma^2}$. If the first (second) term on RHS of (10) after (c) is always smaller than the second (first) one $\forall \gamma$, i.e. if $\gamma \geq \beta q_u^{U,\epsilon} / (1 + \frac{\beta}{l} q_e^E)$ (versus $\gamma \leq \beta q_u^{U,\epsilon} / (1 + \frac{\beta}{l} q_e^E)$), the maximum of the normalized sum-rate is achieved for the smallest (largest) feasible γ as that function is monotonically decreasing (increasing) in γ . Otherwise, it is achieved where they intersect. Hence, for all cases the best overlap rate is attained if γ satisfies (9). \square

OPC's dependence on the joint overlapping characteristics make it challenging to use in optimizing scheduling. In the sequel we shall consider the following linear policy which

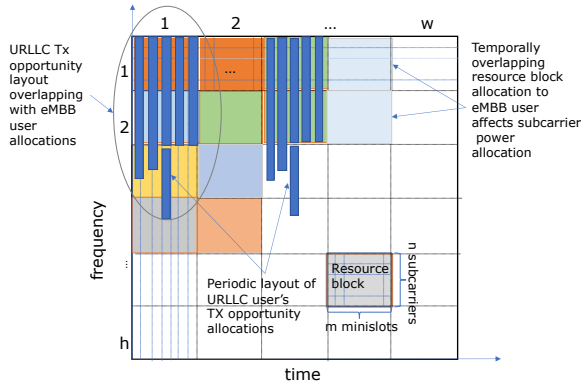


Fig. 1: An uplink frame with optimized layout for semi-persistent URLLC users and eMBB users with dynamic demands opportunistically overlapping RBs over the layout.

approximates OPC in the low SNR per subcarrier regime and/or poor eMBB channel quality, i.e. when $\frac{\beta q_e^E}{l} \ll 1$, and depends only on $q_u^{U,\epsilon}$.

Proposition 2. Linear Power Control (LPC) policy. We refer to a linear eMBB power control policy as setting the target γ in Eq. (1) to

$$\gamma^{LPC}(q_u^{U,\epsilon}, \kappa) = \kappa \beta q_u^{U,\epsilon}, \quad \text{for some } \kappa \in (0, 1). \quad (11)$$

Note that this choice γ^{LPC} ensures the associated eMBB power control derived from the bound in Eq. (1) depends solely on q_e^E .

Granularity of resource allocation. For reference we sketch an uplink frame structure in Fig. 1 illustrating the optimized layout for periodic scheduling of semi-persistent strips of URLLC users and opportunistic scheduling of eMBB users over the URLLC layout. As shown eMBB users are allocated RBs, i.e. blocks of m minislots and n subcarriers while URLLC users are periodically allocated strips of consecutive subcarriers across minislots. Note that an eMBB user allocated RBs in the same time slot (column) will need to spread its power across the associated subcarriers. The efficient overlapping of URLLC user strips over eMBB users' RBs corresponds to a packing problem explored in Section IV of this paper.

IV. JOINT SCHEDULING OF URLLC AND EMBB TRAFFIC

In this section we adopt LPC throughout. Since LPC based eMBB power control results in a URLLC rate that depends solely on the overlapping URLLC channel quality, the resource requirements to meet URLLC users' demands can be determined without prior knowledge of the overlapping eMBB users' channel qualities. This permits one to develop a practical approach to joint URLLC/eMBB scheduling which can be decomposed into two sub-problems: 1) optimizing the layout of URLLC users demands; and 2), scheduling the eMBB users to opportunistically overlap their transmissions across the URLLC layout.

A. Optimizing URLLC layout for semi-persistent demands

We shall consider resource allocation for semi-persistent URLLC uplink user traffic along with dynamic eMBB uplink

demands. A URLLC user $u \in U$ specifies its requirement in terms of periodic transmission opportunities, say b_u bits every p minislots, which are pre-allocated and may or may not be used for transmission. For simplicity URLLC users share the same period p . This corresponds to a request for $k_{u,e}^U(b_u)$ (defined in (4)) consecutive subcarriers every p minislots. Note that under LPC $k_{u,e}^U(b_u)$ is independent of e 's channel quality so henceforth we write $k_u^U(b_u)$. It may be constrained to take a particular set of possible integers, see e.g., [24].

A layout of URLLC users' demands across a frame is defined based on a 0-1 matrix $\mathbf{A} = (A_{x,y}^u : u \in U, x = 1, \dots, f_w, y = 1, \dots, f_h)$. If the assignment of user u 's first periodic requirement to minislot x starting at subcarrier y is such that $y \leq f_h - k_u^U(b) + 1$ then $A_{x,y}^u = A_{x,y+1}^u = \dots = A_{x,y+k_u^U(b)-1}^u = 1$, corresponding to $k_u^U(b)$ consecutive subcarrier assignments, and 0 otherwise. Also to ensure feasibility, sub-carrier assignments must be such that $\forall x, y$ we have that $\sum_{u \in U} A_{x,y}^u \leq 1$, i.e., the URLLC users' transmission opportunities do not overlap.

Let $R^{r,s} \subset F$ denote the minislots and subcarrier indices corresponding to RB $(r, s) \in \{1, 2, \dots, w\} \times \{1, 2, \dots, h\}$, i.e., $R^{r,s} = \{(r-1)m+1, \dots, rm-1\} \times \{(s-1)n+1, \dots, sn-1\}$. Under a given layout \mathbf{A} user u 's assignment overlaps with that of RB (r, s) if for some $(x, y) \in R^{r,s}$ we have $A_{x,y} = 1$, we shall let $U_{r,s}(\mathbf{A})$ denote the subset of users whose assignments overlap with RB (r, s) . Fig. 1 exhibits an example URLLC layout over an OFDMA frame.

Under LPC eMBB users' transmit powers are independent of the channel quality of the URLLC user(s) it may overlap, i.e., it only depends on whether there is an overlap. Thus to maximize eMBB user capacity one need only focus on constructing URLLC layouts \mathbf{A} which maximize $F(\mathbf{A})$, i.e., the number of RBs with no URLLC overlaps in the layout \mathbf{A} .

The maximization of $F(\mathbf{A})$ can be connected to makespan minimization in multi-processor scheduling. Let us consider the problem of placing the first cluster of periodic demands (see Fig. 1) in a URLLC layout on an infinite column of RBs with width m minislots, i.e., $w = 1$ and $h = \infty$. This can be viewed as placing the demands on m parallel servers. An offline algorithm for placing such demands in a *greedy first fit decreasing* manner would order demands in decreasing size, and place them sequentially on the least loaded server. The makespan which corresponds to the sum of the loads on the most loaded server achieved by this algorithm is known to be within $\frac{4}{3}$ of the optimal that is achievable [26]. Therefore, minimizing the makespan is the same as maximizing the number of free RBs $F(\mathbf{A})$. We can extend this idea to the case with finite height h and $w > 1$ by allocating resources in a new RB column of width m minislots adjacent to the current column if a URLLC user's demand no longer fits in the current column. Once we have allocated resources to URLLC users, we then repeat this allocation every p minislots. The algorithm is formalized below and an example of the URLLC layout created using this algorithm is shown in Fig. 1.

Greedy First Fit Decreasing. Let U' be the set of URLLC users which have not yet been allocated resources. Initially

$U' = U$. Every p minislots we implement the following algorithm:

- 1) Let \tilde{u} be the URLLC user with highest value of $k_u^U(b_u)$ in U' .
- 2) Find the minislot (\tilde{m}) with least of number of RBs overlapped by URLLC users.
- 3) If \tilde{u} fits in the current column, then place user \tilde{u} on minislot \tilde{m} and set $U' \leftarrow U' \setminus \tilde{u}$. Otherwise move to a new adjacent column of width m slots.
- 4) Go to step 1 until $U' = \emptyset$.

B. Opportunistic Overlapping: eMBB scheduling over URLLC Layouts

Recall URLLC users are assumed to be persistent. Thus, once a URLLC layout \mathbf{A} is determined it is fixed for a period of time. This induces a distribution for each eMBB user's rate on each RB depending on its time varying channel quality and whether it overlaps with the URLLC layout, i.e., (7) and (8).

eMBB power control policy. There are two factors which limit the eMBB users' transmit power per subcarrier. First, the need to protect overlapping URLLC users' transmissions captured by (1). Second, the fact that p_{\max} will be divided equally among the subcarriers (RBs in same column) allocated to the eMBB user in a given slot. For simplicity we shall make the following assumption to ensure that eMBB transmit power is limited to p_{\max} .

Assumption 1. Each eMBB user will be allocated at most ζ RBs in a slot. The eMBB user's transmit power per sub-carrier is $\frac{p_{\max}}{\zeta n}$ and if it overlaps with a URLLC user the minimum of $\frac{p_{\max}}{\zeta n}$ and the bound of (2) is applied under LPC.

Wireless channel variations. We shall assume that eMBB users' channel qualities are fixed for the duration of a frame and i.i.d. across frames. The marginal distribution for this process is modeled by a random variable Z taking values in a finite set of system states $\mathcal{Z} = \{1, \dots, |\mathcal{Z}|\}$ with probability mass function $p_{\mathcal{Z}}(\cdot)$. We shall denote the rate achieved by an eMBB user e in RB (r, s) under URLLC layout \mathbf{A} and channel state z is given by $c_{\mathbf{A},e}^{E,z}(r, s)$. The scheduler is assumed to know $c_{\mathbf{A},e}^{E,z}(r, s)$ for all e, z , and (r, s) , and can thus opportunistically exploit such variations in allocating resources to eMBB users.

Capacity set for eMBB traffic. It should be clear that the capacity region of eMBB users depends on the URLLC layout \mathbf{A} and the power control policy induced by Assumption 1. In order to capture the capacity set $\mathcal{C}_{\mathbf{A}} \subset \mathbb{R}_+^{|\mathcal{E}|}$ we define the set of all feasible allocation vectors as $\mathcal{B} \subset [0, 1]^{w \times h \times |\mathcal{E}| \times |\mathcal{Z}|}$. For each allocation vector $\mathbf{B} \in \mathcal{B}$, $\mathbf{B}(r, s, e, z)$ is the probability that RB (r, s) is allocated to eMBB user e in state z . One may also view $\mathbf{B}(r, s, e, z)$ as the fraction of time that we allocate RB (r, s) to eMBB user e when network state is z . Therefore, for all $r \in [1 : w]$, $s \in [1 : h]$ and $z \in [1 : |\mathcal{Z}|]$ we have the constraint $\sum_{e=1}^{|\mathcal{E}|} \mathbf{B}(r, s, e, z) \leq 1$.

We define a Stationary Scheduling Policy π as a mapping from (\mathbf{A}, z) to \mathcal{B} , i.e., given a URLLC layout and network

state z , π is associated with allocation vector \mathbf{B}^π . Let $\Pi_{\mathbf{A}}$ be the set of all Stationary Scheduling Policies for a given URLLC layout \mathbf{A} . We define the capacity set $\mathcal{C}_{\mathbf{A}} \subset \mathbb{R}_+^{|\mathcal{E}|}$ for eMBB traffic as the set of long term rates achievable under policies in Π . Let $\mathbf{c}^\pi = (c_e^\pi | e \in \mathcal{E})$ where

$$c_e^\pi = \sum_{z \in \mathcal{Z}} p_{\mathcal{Z}}(z) \left(\sum_{r=1}^w \sum_{s=1}^h \mathbf{B}^\pi(r, s, e, z) c_{\mathbf{A},e}^{E,z}(r, s) \right). \quad (12)$$

Then the capacity region is given by $\mathcal{C}_{\mathbf{A}} = \{\mathbf{c} \in \mathbb{R}_+^{|\mathcal{E}|} | \exists \pi \in \Pi_{\mathbf{A}} \text{ such that } \mathbf{c} \leq \mathbf{c}^\pi\}$. Note that this capacity region depends on the distributions of the channel states, URLLC layout and eMBB power control policy.

Utility maximization problem. We shall assume eMBB users have infinitely backlogged uplink queues. Consider a standard sum utility maximization framework where each eMBB user e has an associated utility function $f_e(\cdot)$ which is a strictly concave, continuous and differentiable function of the average rate c_e experienced by the eMBB user. Our objective is to determine a scheduling policy π^* which attains the maximum in the following optimization problem:

$$\max_{\pi} \left\{ \sum_{e \in \mathcal{E}} f_e(c_e^\pi) | \mathbf{c}^\pi \in \mathcal{C}_{\mathbf{A}} \right\}. \quad (13)$$

Determining an optimal policy requires knowledge of wireless channel variation statistics. A standard approach [27] to circumvent this issue is to use a gradient based online scheduling algorithm which depends on the average rate based on past allocations to eMBB users $\{\bar{\tau}_e | e \in \mathcal{E}\}$, marginal utilities of users $\left\{ f'_e(\bar{\tau}_e) := \left. \frac{df_e(x)}{dx} \right|_{x=\bar{\tau}_e} | e \in \mathcal{E} \right\}$ and current channel rates $\left\{ c_{\mathbf{A},e}^{E,z}(r, s) | e \in \mathcal{E} \right\}$. We first allocate RBs to eMBB users on a given slot, i.e., a fixed value of r . Once we allocate all RBs in a slot, we then move onto the next slot and so on. Let $E_r(s) \subset \mathcal{E}$ be the set of eMBB users which have been allocated less than ζ RBs after performing the allocation to RB (r, s) . For the RB $(r, s + 1)$, we schedule an eMBB user e^* such that

$$e^* \in \operatorname{argmax}_{e \in E_r(s)} \left\{ c_{\mathbf{A},e}^{E,z}(r, s) f'_e(\bar{\tau}_e) \right\}. \quad (14)$$

After allocation to each RB, we update $\bar{\tau}_e$ as follows:

$$\bar{\tau}_e \leftarrow (1 - \alpha)\bar{\tau}_e + \alpha(1 - \epsilon)c_{\mathbf{A},e}^{E,z}(r, s), \quad (15)$$

where $\alpha > 0$. We continue this process until all RBs in a frame are exhausted. The following theorem summarizes the optimality of the above scheduler.

Theorem 1. Given a fixed URLLC layout and LPC, the channel and URLLC layout aware utility maximizing scheduler described by (14) and (15) achieves the optimal eMBB sum utility as $\alpha \rightarrow 0$.

The proof of the above theorem is more or less standard at this point, see e.g., [27]. The only difference with respect to previous work is that the fixed URLLC layout influences the possible achievable rates under different channel states.

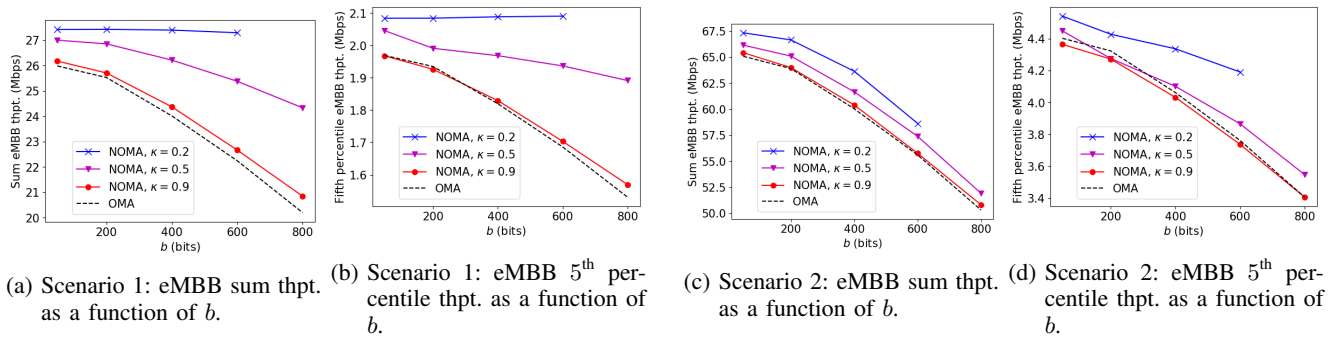
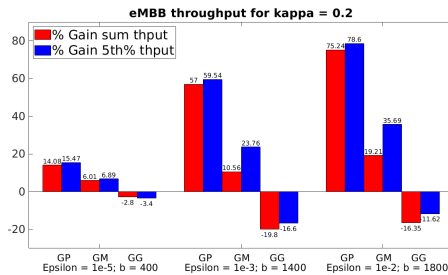


Fig. 2: Simulation results for Scenarios 1 and 2.


 Fig. 3: Comparison of eMBB sum and 5th percentile thpt.

C. Performance Evaluation

We consider an uplink OFDMA system with 10 eMBB users and 10 URLLC users. We follow the OFDMA numerology in [28] with carrier frequency 900 MHz, system bandwidth 40 MHz and sub-carrier bandwidth 15 KHz. Each slot is of duration 1 msec. and has 14 OFDMA symbols. A slot consists of 7 minislots with a minislot consisting of 2 OFDMA symbols. An RB for eMBB user consists of 12 subcarriers and 14 OFDMA symbols. For URLLC users resource allocation is performed at a granularity of 12 subcarriers and 2 OFDMA symbols. We choose $p_{\max} = 23$ dBm and thermal noise power $N_oW = -104$ dBm based on the simulation parameters mentioned in [29]. The channel gain consists of path loss and small-scale fading. The path loss in dB at a distance d meters is given by the expression $120.9 + 37.6 \log_{10}(d/1000)$. For convenience, we fix the inter-cell interference at -80 dBm and we choose $\zeta = 30$, $f_e(\cdot) = \log(\cdot)$, and $p = 1$.

We classify users based on their distance to the BS into three different classes; users within a distance of 30 meters from the BS have *good* channel conditions, users between 30 and 60 meters have *medium* quality channel conditions and users at a distance greater than 60 meters have *poor* channel conditions, i.e., they are the cell edge.

eMBB throughput as a function of κ and b : There are two factors which determine the eMBB sum/5th percentile throughput 1) eMBB capacity in overlapped RBs vs non-overlapped RBs; and 2), the number of RBs required by URLLC users. Observe that when we reduce κ , URLLC capacity per sub-carrier decreases but eMBB users get higher capacity in overlapping RBs (see (11)). However, for a given b

reducing κ increases the URLLC bandwidth requirement and hence reduces the free RBs for eMBB users. Reducing κ even further for a given value of b results in a scenario where we cannot fit URLLC demands in 40 MHz bandwidth. Hence, the performance of NOMA w.r.t. OMA depends on the balance between the above factors.

We consider two simulation scenarios in Figure 2 with $\epsilon = 10^{-5}$ and compare the performance of NOMA w.r.t. OMA for different values of κ resulting in different power control limitation for eMBB traffic based on (11). In both, URLLC users have good channel conditions. eMBB users have *poor* and *medium* channel conditions in Scenarios 1 and 2, respectively. For NOMA we use the URLLC layout generation strategy and eMBB scheduling strategy in Sec. IV-A and Sec. IV-B, respectively. In OMA, we create URLLC layout based on OMA URLLC capacity expression (5). However, unlike NOMA, in OMA we schedule eMBB users only in RBs which are not overlapped with URLLC users. We average the results over 30 random drops of eMBB/URLLC users.

In Scenario 1, since eMBB users have *poor* channels on average, hence, for low values of κ , eMBB users at the edge can transmit with little or no power back-off. Hence, for $\kappa = 0.2$, eMBB throughput is almost insensitive to URLLC load (see Figures 2a and 2b). However, we can support URLLC load of only up to $b = 600$ bits with $\kappa = 0.2$. With higher values of κ , we can support higher values of b and also the eMBB throughput is more sensitive to URLLC loads. For $b = 800$ bits with $\kappa = 0.5$, with NOMA there is approximately 19.5% gain in sum throughput and 26.6% gain in the 5th percentile throughput with respect to OMA. With OMA, eMBB sum throughput decreases almost linearly as a function of b . In Scenario 2 since eMBB users have better channels than Scenario 1, the power back-offs required are higher. Therefore, the gains w.r.t. OMA diminish to less than 5% (see Figures 2c and 2d). Hence, we observe higher gains for NOMA when eMBB users' channel conditions are much poorer than URLLC users' channel conditions.

eMBB throughput as a function of ϵ : In Fig. 3, we compare the percentage gains in sum throughput and 5th percentile throughput w.r.t. OMA for different ϵ and three different scenarios. For example, the scenario with label GP implies that URLLC users have good (G) channel conditions and eMBB

users have poor (P) channel conditions. Similarly, we have named other scenarios. Increasing ϵ improves URLLC users' reliable channel quality $q_u^{U,\epsilon}$. Hence, we can support a higher value of b for a given system bandwidth. However, for OMA to support high values of b , a large portion of the bandwidth must be reserved for URLLC, reducing the overall eMBB throughput. Hence, we see higher gains with increasing ϵ , for example, with $\kappa = 0.2$ we see 14.08% gain in eMBB sum throughput for $b = 400$ and $\epsilon = 1e-5$ which increases to 75.24% at $b = 1800$ and $\epsilon = 1e-2$. Further we observe that in scenarios with good channels for both URLLC and eMBB users, OMA performs better than NOMA.

V. CONCLUSION AND FUTURE WORK

We have presented a detailed framework, analysis and performance evaluation addressing URLLC/eMBB overlapping and scheduling on uplink cellular resources leveraging a NOMA based physical layer framework. We proposed the concept of transmission opportunities for URLLC users to meet their latency constraints. We derived sum-rate optimal power control and proposed a simple approximation to it. One observation, perhaps not unexpected, is that the potential improvements in efficiency, depend in a complex manner on joint power control strategies and on the spatial user loads of URLLC/eMBB traffic the system supports. Our particular focus was on achieving improved efficiency in supporting the low delay/high reliability of URLLC traffic through scheduling overlapping eMBB traffic, though additional gains would be possible through overlapping eMBB user traffic with each other, see e.g., [30]. In future work we will improve upon our framework by considering additional trade-offs embedded in the joint power control policy that would lead to better optimization of overlapping transmissions.

REFERENCES

- [1] R. Srikant and L. Ying, *Communication networks: An optimization, control, and stochastic networks perspective*. Cambridge University Press, 2013.
- [2] Y. Ozcan and C. Rosenberg, "Uplink scheduling in multi-cell OFDMA networks: A comprehensive study," *IEEE Transactions on Mobile Computing*, May 2020.
- [3] A. Hussein, C. Rosenberg, and P. Mitran, "Proportional fair downlink NOMA; from a centralized analysis to practical schemes." [Online]. Available: <https://eecs.iisc.ac.in/EECS2020/details/talk.php?id=16>
- [4] T. Pijnappel, S. Borst, and P. Whiting, "Joint scheduling of low-latency and best-effort flows in 5G wireless networks," in *Proc., IEEE International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks*, Jun. 2020, pp. 1–8.
- [5] A. Anand, G. De Veciana, and S. Shakkottai, "Joint scheduling of URLLC and eMBB traffic in 5G wireless networks," *IEEE/ACM Transactions on Networking*, vol. 28, no. 2, pp. 477–490, Feb. 2020.
- [6] C. Tang, X. Chen, Y. Chen, and Z. Li, "Dynamic resource optimization based on flexible numerology and markov decision process for heterogeneous services," in *Proc., IEEE International Conference on Parallel and Distributed Systems*, Dec. 2019, pp. 610–617.
- [7] G. Pocovi, K. I. Pedersen, and P. Mogensen, "Joint link adaptation and scheduling for 5G ultra-reliable low-latency communications," *IEEE Access*, vol. 6, pp. 28912–28922, May 2018.
- [8] A. Pradhan and S. Das, "Joint preference metric for efficient resource allocation in co-existence of eMBB and URLLC," in *Proc., International Conference on Communication Systems & NETWORKS*, Jan. 2020, pp. 897–899.
- [9] G. Pocovi, K. I. Pedersen, and P. Mogensen, "Multiplexing of latency-critical communication and mobile broadband on a shared channel," in *Proc., IEEE Wireless Communications and Networking Conference*, Apr. 2018, pp. 1–6.
- [10] P. Korrai, E. Lagunas, S. K. Sharma, S. Chatzinotas, A. Bandi, and B. Ottersten, "A RAN resource slicing mechanism for multiplexing of eMBB and URLLC services in OFDMA based 5G wireless networks," *IEEE Access*, vol. 8, pp. 45674–45688, Mar. 2020.
- [11] K. Zhang, X. Xu, J. Zhang, B. Zhang, X. Tao, and Y. Zhang, "Dynamic multicommunity based joint scheduling of eMBB and uRLLC in 5G networks," *IEEE Systems Journal*, Apr. 2020.
- [12] N. Ksairi and M. Kountouris, "Timely scheduling of URLLC packets using precoder compatibility estimates," in *Proc., IEEE Annual International Symposium on Personal, Indoor and Mobile Radio Communications*, Sep. 2019, pp. 1–6.
- [13] A. A. Esswie, K. I. Pedersen, and P. E. Mogensen, "Preemption-aware rank offloading scheduling for latency critical communications in 5G networks," in *Proc., IEEE Vehicular Technology Conference (VTC-Spring)*, Apr. 2019, pp. 1–6.
- [14] 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 Workshops*, Dec. 2018, pp. 1–6.
- [15] —, "Capacity optimization of spatial preemptive scheduling for joint URLLC-eMBB traffic in 5G new radio," in *Proc., IEEE Globecom Workshops (GC Wkshps)*, Dec. 2018, pp. 1–6.
- [16] —, "Opportunistic spatial preemptive scheduling for URLLC and eMBB coexistence in multi-user 5G networks," *IEEE Access*, vol. 6, pp. 38451–38463, Jul. 2018.
- [17] Z. Yang, Z. Ding, P. Fan, and N. Al-Dhahir, "A general power allocation scheme to guarantee quality of service in downlink and uplink NOMA systems," *IEEE Transactions on Wireless Communications*, vol. 15, no. 11, pp. 7244–7257, Aug. 2016.
- [18] Z. Ding, M. Peng, and H. V. Poor, "Cooperative non-orthogonal multiple access in 5G systems," *IEEE Communications Letters*, vol. 19, no. 8, pp. 1462–1465, Jun. 2015.
- [19] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5G nonorthogonal multiple-access downlink transmissions," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 8, pp. 6010–6023, Sep. 2015.
- [20] U. Challita, K. Hiltunen, and M. Tercero, "Performance evaluation for the co-existence of eMBB and URLLC networks: Synchronized versus unsynchronized TDD," in *Proc., IEEE Vehicular Technology Conference (VTC-Fall)*, Sep. 2019, pp. 1–6.
- [21] R. Abreu, T. Jacobsen, K. Pedersen, G. Berardinelli, and P. Mogensen, "System level analysis of eMBB and grant-free URLLC multiplexing in uplink," in *Proc., IEEE Vehicular Technology Conference (VTC-Spring)*, Apr. 2019, pp. 1–5.
- [22] T. H. Jacobsen, R. Abreu, G. Berardinelli, K. I. Pedersen, I. Z. Kovács, and P. Mogensen, "Multi-cell reception for uplink grant-free ultra-reliable low-latency communications," *IEEE Access*, vol. 7, pp. 80208–80218, Jun. 2019.
- [23] C. Zhang, Y. Liu, Z. Qin, and Z. Ding, "Semi-grant-free NOMA: A stochastic geometry model," *arXiv preprint arXiv:2006.13286*, Jun. 2020.
- [24] "Evolved universal terrestrial radio access (EUTRA); Physical layer procedures," 3GPP TS36.213; V15.10.0.
- [25] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2307–2359, Apr. 2010.
- [26] R. L. Graham, "Bounds on multiprocessing timing anomalies," *SIAM Journal on Applied Mathematics*, vol. 17, no. 2, pp. 416–429, Mar. 1969. [Online]. Available: <http://www.jstor.org/stable/2099572>
- [27] A. L. Stolyar, "On the asymptotic optimality of the gradient scheduling algorithm for multiuser throughput allocation," *Operations research*, vol. 53, no. 1, pp. 12–25, Feb. 2005.
- [28] A. Gosh, "5G New Radio (NR): Physical layer overview and performance," in *Proc. IEEE Comm. Theory Workshop*, May 2018.
- [29] "LTE; evolved universal terrestrial radio access (E-UTRA); radio frequency (RF) system scenarios." [Online]. Available: https://www.etsi.org/deliver/etsi_tr/136900_136999/136942/10.02.00_60/tr_136942v100200p.pdf
- [30] Y. Polyanskiy, "Information-theoretic perspective on massive multiple-access (tutorial)," North-American School of Information Theory, Texas A&M University, College Station, TX, May 2018.