Stochastic optimization of URLLC-eMBB joint scheduling with queuing mechanism

PLEASE CITE THE PUBLISHED VERSION

https://doi.org/10.1109/lwc.2020.3046628

PUBLISHER

Institute of Electrical and Electronics Engineers (IEEE)

VERSION

AM (Accepted Manuscript)

REPOSITORY RECORD

Zhang, Wenheng, Mahsa Derakhshani, and Sangarapillai Lambotharan. 2020. “Stochastic Optimization of Urllc-embb Joint Scheduling with Queuing Mechanism”. Loughborough University. https://hdl.handle.net/2134/13517087.v1.

Stochastic Optimization of URLLC-eMBB Joint Scheduling with Queuing Mechanism

Wenheng Zhang, Mahsa Derakhshani, and Sangarapillai Lambotharan

***Abstract*—This paper proposes a dynamic joint scheduling for the ultra-reliable low-latency communication (URLLC) and enhanced mobile broadband (eMBB) traffic at a sub-frame level with a queuing mechanism, which monitors and controls the latency of each URLLC packet in real time to ensure its strict requirements. We analytically derive the outage probability (i.e., the probability of any URLLC packet drop over all transmission channels) and URLLC expected throughput in addition to the expected value of served URLLC packets. Then, a stochastic optimization problem is formulated to maximize the total throughput for the eMBB services, constraining the URLLC outage probability. Numerical results confirm effectiveness of our queuing policy to significantly reduce the URLLC loss rate while ensuring that the total eMBB throughput is not affected.**

***Index Terms*—URLLC, eMBB, punctured scheduling, queuing.**

I. Introduction

ENHANCED Mobile Broadband (eMBB) and ultra-reliable and low-latency communications (URLLC) are two main services that will be supported by flexible network operation in future wireless communication networks. The URLLC service targets mission critical applications [1]–[5]. The challenge of URLLC is an internal trade-off between its latency and reliability requirements. Besides, the reliability of URLLC is influenced by the coexistence with other types of traffic. Thus, channel scheduling faces a critical challenge on simultaneously implementing these services with different targets. 5G services require a high peak data rate for eMBB users up to gigabits per second, while the URLLC traffic transmits small payload within 1ms latency and at least 99.999% success probability [2]. In this case, the BS should provide stable and satisfactory transmission channels for sporadic transmission of URLLC packets but also maximize the eMBB throughput.

Dynamic Scheduling of eMBB and URLLC has received attention recently [6]–[11]. [6] proposed the idea of scheduling eMBB and URLLC packets on different time scales. Subsequently, the idea of punctured scheduling, i.e., suspension of the eMBB transmission when a URLLC packet is decided to be served, was proposed by [7]. This scheme has been further explored by [8] that studied the joint scheduling optimization investigating different loss functions (i.e. linear, convex and threshold models) to model the effects of punctured URLLC transmission on eMBB throughput. Moreover, [9] analyzed the eMBB throughput and operational complexity in a dynamic mode. With a risk-aware framework, [10] proposed a risk-sensitive optimization which punctures more URLLC packets on the robust set of eMBB users to protect the sensitive group. However, in these frameworks, the incoming URLLC packets are either served immediately or dropped on each mini slot. This incurs significant URLLC packet drops, which could be avoided to some extent with an appropriate queuing mechanism. Furthermore, further researches are needed to focus on the time-varying channels.

Addressing these existing challenges, in this letter, we propose a stochastic scheduling scheme for URLLC and eMBB coexistence considering a $M/G/\infty$ queuing mechanism for sporadic URLLC packets, this helps the system to meet the URLLC latency requirement by queuing each packet in the line up to two mini slots. Comparing with the immediate drop policy of previous works (e.g., [8]), our scheme improves the URLLC reliability within the advised latency by 3GPP [2].

Moreover, an outage-constrained stochastic optimization problem is formulated under assumption of time-varying channels. The aim is to maximize the total throughput of eMBB users on different CSI while restricting the URLLC outage probability (i.e., the probability of any URLLC drops) and URLLC expected throughput. We use successive convex approximation (SCA) [12] to convert the original non-convex optimization into linear integer programming for eMBB scheduling and geometric programming (GP) for URLLC placement scheduling. We also built the eMBB-URLLC coexistence system of [8] (i.e. without applying URLLC queuing system) and compared the optimization results in terms of mean value of total eMBB throughput and URLLC outage probability.

II. System Design

This paper considers a downlink dynamic scheduling of an eMBB-URLLC coexistence system. We assume a multiple-channel scenario and study scheduling in different frequency channels and time intervals as shown in Fig. 1. Each eMBB slot of one millisecond contains M URLLC mini slots and the set is represented by $\mathcal{M}$. The set of frequency channels is $\mathcal{B} = \{1, 2, \ldots, B\}$. The set of eMBB users is $\mathcal{U} = \{1, 2, \ldots, U\}$. The set of possible channel states on each eMBB slot is denoted by $\mathcal{C} = \{1, 2, \ldots, C\}$. The CSI of all eMBB users on different channels at different time slot is represented by a $B \times U$ matrix, i.e., $\boldsymbol{s}_{\mathrm{e}} = [s^{b}_{\mathrm{e},u}]$, where $s^{b}_{\mathrm{e},u} \in \mathcal{C}$, $\forall u \in \mathcal{U}$, $\forall b \in \mathcal{B}$. We assume the CSI of URLLC user on different channels is represented by a $B \times 1$ matrix, $\boldsymbol{s}_{\mathrm{r}}$. We use $\boldsymbol{s} = [\boldsymbol{s}_{\mathrm{r}}, \boldsymbol{s}_{\mathrm{e}}]$ to denote the CSI of these two kinds of traffic, the probability mass function (PMF) of $\boldsymbol{s}$ is $\mathrm{P}_{S}(\boldsymbol{s})$, $\forall \boldsymbol{s} \in \mathcal{S}$, where the cardinality of $\mathcal{S}$ is $S = C^{(B+1)U}$.

At each time slot, the BS decides the allocation for all eMBB users on different channels. The eMBB allocations is presented by a $B \times S \times U$ matrix $\boldsymbol{I} = [I^{\boldsymbol{s}}_{u,b}]$, where $I^{\boldsymbol{s}}_{u,b} \in \{0, 1\}$ is a binary variable declaring whether channel b is allocated to eMBB user u or not when the CSI is $\boldsymbol{s}$. It is assumed that each channel is always fully occupied by one of the eMBB users

This work was supported by the Royal Academy of Engineering under the Leverhulme Trust Research Fellowship scheme (Derakhshani-LTRF1920\16\67). W. Zhang, M. Derakhshani, and S. Lambotharan are with the Wolfson School of Mechanical, Electrical and Manufacturing Engineering, Loughborough University, UK (email: W.Zhang2, M.Derakhshani, S.Lambotharan@lboro.ac.uk).

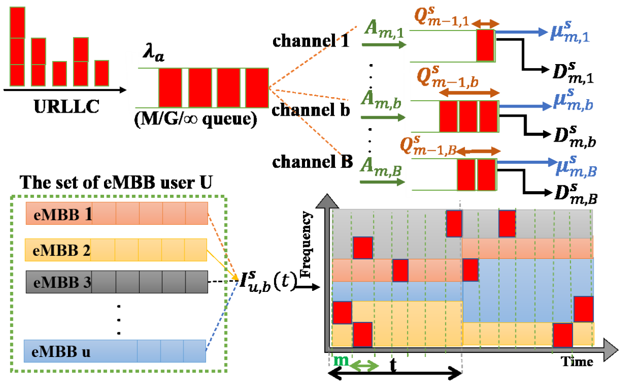

Fig. 1. Scheduling policy for the eMBB-URLLC joint system.

on each time slot, i.e. $\sum_{u\in\mathcal{U}} I^{\boldsymbol{s}}_{u,b} = 1$, $\forall \boldsymbol{s} \in \mathcal{S}$, $\forall b \in \mathcal{B}$. Once the BS receives an incoming URLLC packet and decides to transmit it on channel b, it suspends the eMBB communication and punctures URLLC on channel b during next mini slot. We design and characterize a queuing mechanism to stochastically restrict the number of dropped URLLC packets which will impact the reliability. Let A_m denote the number of URLLC packet arrivals between mini slot (*m-1*) and mini slot *m*. We assume Poisson distribution for URLLC packet arrivals with a mean rate λ_a. The arriving packets are divided between B different channels uniformly as shown Fig. 1. To ensure that the delay requirement of URLLC served packets is always complied with the 3GPP standard (i.e., 0.25-0.3 milliseconds [2]) while reducing the number of packet drops, it is allowed each URLLC packet to wait in the queue up to two mini slots in the proposed framework. At the beginning of each mini slot, BS decides the transmission of URLLC packets. If a URLLC packet is already queued for one mini slot and will not be served at next mini slot, it will be dropped.

We define $w^{\boldsymbol{s}}_{u,b}$ as the weight of puncturing the eMBB user u on channel b when CSI is **s**, $w^{\boldsymbol{s}}_{u,b} \in [0,1], \forall \boldsymbol{s} \in \mathcal{S}, \forall b \in \mathcal{B}, \forall u \in \mathcal{U}$. Based on the value of $w^{\boldsymbol{s}}_{u,b}$, at each mini slot, function $b(w^{\boldsymbol{s}}_{u,b})$ applies Bernoulli random variable generator to obtain a URLLC transmission decision for each channel. $b(w^{\boldsymbol{s}}_{u,b}) = 1$ indicates that a URLLC packet will be punctured during eMBB u on channel b transmission when CSI is **s** (this occurs with probability $w^{\boldsymbol{s}}_{u,b}$), and 0 indicates that the BS continues the transmission of eMBB user signal on that mini slot (this occurs with probability $1 - w^{\boldsymbol{s}}_{u,b}$). If $I^{\boldsymbol{s}}_{u,b} = 1$, BS will operate $b(w^{\boldsymbol{s}}_{u,b})$ for each mini-time slot on channel b. Accordingly, $C^{\boldsymbol{s}}_{m,b} \in \{0,1\}$ is a random number realizing if the mini slot *m* will be allocated for URLLC transmission or not on channel b if CSI is **s** and can be represented as

$$C^{\boldsymbol{s}}_{m,b} = \sum\nolimits_{u\in\mathcal{U}} I^{\boldsymbol{s}}_{u,b} b(w^{\boldsymbol{s}}_{u,b}), \quad \forall m \in \mathcal{M}, \forall b \in \mathcal{B} \tag{1}$$

Subsequently, the number of served URLLC packets at mini slot *m* on channel b when CSI is **s** can be calculated as

$$\mu^{\boldsymbol{s}}_{m,b} = \min(C^{\boldsymbol{s}}_{m,b}, Q^{\boldsymbol{s}}_{m-1,b} + A_{m,b}), \tag{2}$$

where $Q^{\boldsymbol{s}}_{m-1,b}$ denotes the number of queued packet at mini time-slot $(m-1)$. Consequently, the number of dropped packets at mini slot m can be computed as

$$D^{\boldsymbol{s}}_{m,b} = \max(Q^{\boldsymbol{s}}_{m-1,b} - \mu^{\boldsymbol{s}}_{m,b}, 0). \tag{3}$$

Based on (1)-(3), when CSI is **s** in channel b, the queuing equation at mini slot m can be represented as

$$Q^{\boldsymbol{s}}_{m,b} = Q^{\boldsymbol{s}}_{m-1,b} + A_{m,b} - (\mu^{\boldsymbol{s}}_{m,b} + D^{\boldsymbol{s}}_{m,b}) \tag{4}$$

Studying the queue dynamics (see Appendix A), with $\lambda = \lambda_a/B$, the expected number of served URLLC packets is a function of eMBB allocations $\boldsymbol{I}$ and puncturing weights $\boldsymbol{W}$ as

$$E[\mu^{\boldsymbol{s}}_m] = \sum\nolimits_{b\in\mathcal{B}} \sum\nolimits_{u\in\mathcal{U}} I^{\boldsymbol{s}}_{u,b} (w^{\boldsymbol{s}}_{u,b})^2 \lambda (w^{\boldsymbol{s}}_{u,b}\lambda - e^{-\lambda})^{-1} \tag{5}$$

III. Dynamic Scheduling

A. Optimization Problem

Here, we formulate a stochastic optimization problem aiming to optimally determine the URLLC puncturing weights $\boldsymbol{W}$ and resource allocation factors $\boldsymbol{I}$ for each CSI **s** over all channels in order to schedule the incoming URLLC packets while maximizing the total throughput of eMBB users. The total expected throughput of eMBB users can be calculated as

$$T_{\text{embb}} = \sum_{b\in\mathcal{B}} \sum_{u\in\mathcal{U}} \sum_{\boldsymbol{s}\in\mathcal{S}} \mathrm{P}_s(\boldsymbol{s}) I^{\boldsymbol{s}}_{u,b} R^{\boldsymbol{s}_{\text{e}}}_{u,b} \Big(1 - w^{\boldsymbol{s}}_{u,b} E\big[\mu^{\boldsymbol{s}}_m\big]\Big) \tag{6}$$

where $R^{\boldsymbol{s}_{\text{e}}}_{u,b} = B_W \log(1 + d_u^{-\kappa} \boldsymbol{s}_{\text{e}} \rho)$ is the peak rate of eMBB user u on channel b where d_u is the distance between the eMBB user u and BS, κ denotes the path loss exponent, B_W is the channel bandwidth and ρ is the transmitted SNR of each eMBB user. In (6), $w^{\boldsymbol{s}}_{u,b} E\big[\mu^{\boldsymbol{s}}_m\big]$ represents the ratio of expected number of mini-slots of eMBB user *u* punctured for URLLC service on channel b to the total number of mini-slots in one eMBB slot under CSI **s** under all channels.

Similarly, the expected throughput of URLLC

$$T_{\text{urllc}} = \sum_{b\in\mathcal{B}} \sum_{u\in\mathcal{U}} \sum_{\boldsymbol{s}\in\mathcal{S}} \mathrm{P}_s(\boldsymbol{s}) I^{\boldsymbol{s}}_{u,b} w^{\boldsymbol{s}}_{u,b} R^{\boldsymbol{s}_{\text{r}}}_{urllc,b} E\big[\mu^{\boldsymbol{s}}_m\big] \tag{7}$$

where $R^{\boldsymbol{s}_{\text{r}}}_{urllc,b} = B_W \log(1 + d_r^{-\kappa} \boldsymbol{s}_{\text{r}} \rho_r) - \frac{1}{\ln 2}\sqrt{\frac{V}{b_l}} Q^{-1}(\xi)$ is the channel rate for URLLC on channel b [13] where b_l is the block-length, ξ is the error probability, d_{r} is the distance between URLLC and the BS, ρ_r denotes the transmission SNR of URLLC, V denotes the so-called channel dispersion, and $Q^{-1}(x)$ is the inverse of the Gaussian Q function.

Based on (6) and (7), the optimization problem can be represented as

$$\mathbb{P}: \max_{\boldsymbol{I},\boldsymbol{W}} \sum_{b\in\mathcal{B}} \sum_{u\in\mathcal{U}} \sum_{\boldsymbol{s}\in\mathcal{S}} \varphi_u \mathrm{P}_s(\boldsymbol{s}) I^{\boldsymbol{s}}_{u,b} R^{\boldsymbol{s}_{\text{e}}}_{u,b} \Big(1 - w^{\boldsymbol{s}}_{u,b} E\big[\mu^{\boldsymbol{s}}_m\big]\Big), \textbf{s.t.,}$$

$$\textbf{C1:} \quad \sum\nolimits_{\boldsymbol{s}\in\mathcal{S}} \mathrm{P}_s(\boldsymbol{s}) \prod\nolimits_{b\in\mathcal{B}} \Pr(D^{\boldsymbol{s}}_{m,b} = 0) \geq \zeta, \forall m \in \mathcal{M}$$

$$\textbf{C2:} \quad \sum\nolimits_{u\in\mathcal{U}} I^{\boldsymbol{s}}_{u,b} = 1, \forall b \in \mathcal{B}, \forall \boldsymbol{s} \in \mathcal{S}$$

$$\textbf{C3:} \quad I^{\boldsymbol{s}}_{u,b} \in \{0,1\}, \forall b \in \mathcal{B}, \forall \boldsymbol{s} \in \mathcal{S}, \forall u \in \mathcal{U}$$

$$\textbf{C4:} \quad w^{\boldsymbol{s}}_{u,b} \in [0,1], \forall b \in \mathcal{B}, \forall \boldsymbol{s} \in \mathcal{S}, \forall u \in \mathcal{U}$$

$$\textbf{C5:} \quad \sum_{b\in\mathcal{B}} \sum_{u\in\mathcal{U}} \sum_{\boldsymbol{s}\in\mathcal{S}} \mathrm{P}_s(\boldsymbol{s}) I^{\boldsymbol{s}}_{u,b} w^{\boldsymbol{s}}_{u,b} R^{\boldsymbol{s}_{\text{r}}}_{urllc,b} E\big[\mu^{\boldsymbol{s}}_m\big] \geq \eta,$$

where the stochastic constraint in **C1** guarantees that the probability of no URLLC drop is always higher than a target threshold ζ (equivalently, probability of any drops is kept limited). **C2** ensures that each channel is always occupied exclusively by one of the eMBB users on each time slot. **C3** and **C4** ensure the values of the eMBB allocation variable $I^{\boldsymbol{s}}_{u,b}$ should be binary and puncturing weight $w^{\boldsymbol{s}}_{u,b}$ should be a decimal for each eMBB user u on channel b and CSI **s**. Also, **C5** guarantees that the expected throughput of URLLC service is no less than a target threshold η. We use φ_u to denote the priority weights of eMBB user u.

Algorithm 1: Dynamic URLLC-eMBB scheduling

```
Variables: $\boldsymbol{I}$, $\boldsymbol{W}$, $\boldsymbol{Y}$, $\boldsymbol{P}$ and $\boldsymbol{X}$
Results: Maximize total throughput of eMBB users
Initialize the puncturing matrix $W_0$;
while Not reaching the maximum iteration number or variables do not converge do
    eMBB allocation scheduling as $\mathbb{P}_1$;
    Initialize the iteration variables;
    while Not reaching the maximum iteration number or variables do not converge do
        URLLC placement scheduling as $\widehat{\mathbb{P}_2}$;
    Update the iteration variables as (8) to (13);
```

B. Proposed Solution

The dynamic scheduling problem is decomposed into two stages. The BS schedules eMBB allocations by fixing the puncturing weights $\boldsymbol{W}$. After updating the eMBB allocation matrix $\boldsymbol{I}$, URLLC packets placement scheduling is optimized.

1) eMBB Scheduling: In this stage, we assume that the URLLC puncturing weight matrix $\boldsymbol{W}$ is fixed when optimizing eMBB allocation matrix $\boldsymbol{I}$. For the eMBB scheduling, we solve the following linear integer programming:

$$\mathbb{P}_1 : \max_{\boldsymbol{I}} \sum_{b\in\mathcal{B}}\sum_{u\in\mathcal{U}}\sum_{\boldsymbol{s}\in\mathcal{S}} \varphi_u \mathrm{P}_s(\boldsymbol{s}) I^{\boldsymbol{s}}_{u,b} R^{\boldsymbol{s}_\mathrm{e}}_{u,b}\Big(1-\frac{\lambda(w^{\boldsymbol{s}}_{u,b})^3}{w^{\boldsymbol{s}}_{u,b}\lambda - e^{-\lambda}}\Big)$$

$$\textbf{s.t.} \qquad \textbf{C2, C3}$$

2) URLLC Placement Scheduling: For optimizing the URLLC puncturing weights, we fix $\boldsymbol{I}$ and then solve $\mathbb{P}$. A new variable matrix is defined as $\boldsymbol{Y} = \left[y^{\boldsymbol{s}}_{u,b}\right]$ where $y^{\boldsymbol{s}}_{u,b} = e^{-\lambda} - w^{\boldsymbol{s}}_{u,b}\lambda$, $\forall b\in\mathcal{B}, \forall\boldsymbol{s}\in\mathcal{S}, \forall u\in\mathcal{U}$. Thus, the optimization problem can be transformed into the complementary geometric programming form (CGP) as,

$$\mathbb{P}_2 : \max_{\boldsymbol{W},\boldsymbol{Y}} \sum_{b\in\mathcal{B}}\sum_{u\in\mathcal{U}}\sum_{\boldsymbol{s}\in\mathcal{S}} \varphi_u \mathrm{P}_s(\boldsymbol{s}) I^{\boldsymbol{s}}_{u,b} R^{\boldsymbol{s}_\mathrm{e}}_{u,b}\Big(1-(w^{\boldsymbol{s}}_{u,b})^2\frac{y^{\boldsymbol{s}}_{u,b}-e^{-\lambda}}{y^{\boldsymbol{s}}_{u,b}}\Big)$$

$$\textbf{s.t.} \qquad \textbf{C1, C4, C5}$$

$$\textbf{C6:} \quad e^{-\lambda}\Big(y^{\boldsymbol{s}}_{u,b}+w^{\boldsymbol{s}}_{u,b}\lambda\Big)^{-1} \le 1, \forall b\in\mathcal{B}, \forall\boldsymbol{s}\in\mathcal{S}, \forall u\in\mathcal{U}$$

We use $p^{\boldsymbol{s}}_b$ to denote the probability of no URLLC packet drop over one time-slot on channel b when CSI is **s**. As shown in Appendix B, $p^{\boldsymbol{s}}_b$ is calculated as

$$p^{\boldsymbol{s}}_b = \sum_{u\in\mathcal{U}} I^{\boldsymbol{s}}_{u,b}\frac{\lambda w^{\boldsymbol{s}}_{u,b} - \Big(\lambda w^{\boldsymbol{s}}_{u,b}\Big)^2\Big(\frac{2+\lambda}{2\lambda}\Big)e^{-\lambda}}{e^{-\lambda}-w^{\boldsymbol{s}}_{u,b}\lambda}, \forall\boldsymbol{s}\in\mathcal{S}, \forall b\in\mathcal{B}$$

Thus, **C1** can be rewritten as $\widetilde{\mathbf{C1}}$,

$$\widetilde{\mathbf{C1}} : \sum\nolimits_{s\in\mathcal{S}} \mathrm{P}_s(\boldsymbol{s})\prod\nolimits_{b\in\mathcal{B}} p^{\boldsymbol{s}}_b \ge \zeta$$

Applying arithmetic-geometric mean approximation (AGMA) (by adding a new constraint **C7**) to change the non-posynomial constraint $\widetilde{\mathbf{C1}}$ to posynomial form. Variable $\mathbf{X} = [x^{\boldsymbol{s}}_{u,b}]$ is introduced to change the non-posynomial objective function $\mathbb{P}_2$ into a posynomial form. $[x^{\boldsymbol{s}}_{u,b}] = (w^{\boldsymbol{s}}_{u,b})^2\big(y^{\boldsymbol{s}}_{u,b} - e^{-\lambda}\big)\big(y^{\boldsymbol{s}}_{u,b}\big)^{-1}$, $\forall b\in\mathcal{B}, \forall\boldsymbol{s}\in\mathcal{S}, \forall u\in\mathcal{U}$. **C8** converts $\boldsymbol{X}$ to the posynomial form. The URLLC scheduling optimization becomes,

$$\widetilde{\mathbb{P}_2} : \min_{\boldsymbol{W},\boldsymbol{Y},\boldsymbol{P},\boldsymbol{X}} \sum_{b\in\mathcal{B}}\sum_{u\in\mathcal{U}}\sum_{\boldsymbol{s}\in\mathcal{S}} \varphi_u \mathrm{P}_s(\boldsymbol{s}) I^{\boldsymbol{s}}_{u,b} R^{\boldsymbol{s}_\mathrm{e}}_{u,b} x^{\boldsymbol{s}}_{u,b} \quad \textbf{s.t.}\widetilde{\mathbf{C1}}, \mathbf{C4}, \mathbf{C5}, \mathbf{C6}$$

$$\mathbf{C7} : \frac{y^{\boldsymbol{s}}_{u,b}(2\lambda^{-1}+1)e^{-\lambda}}{p^{\boldsymbol{s}}_b y^{\boldsymbol{s}}_{u,b}e^{\lambda} + y^{\boldsymbol{s}}_{u,b}e^{\lambda} + e^{-2\lambda}\Big(\frac{2+\lambda}{2\lambda}\Big) + \big(y^{\boldsymbol{s}}_{u,b}\big)^2\Big(\frac{2+\lambda}{2\lambda}\Big)} \le 1,$$
$$\forall b\in\mathcal{B}, \forall\boldsymbol{s}\in\mathcal{S}, \forall u\in\mathcal{U}$$

$$\mathbf{C8} : \qquad \frac{y^{\boldsymbol{s}}_{u,b}{w^{\boldsymbol{s}}_{u,b}}^2}{y^{\boldsymbol{s}}_{u,b}x^{\boldsymbol{s}}_{u,b} + e^{-\lambda}{w^{\boldsymbol{s}}_{u,b}}^2} \le 1, \forall b\in\mathcal{B}, \forall\boldsymbol{s}\in\mathcal{S}, \forall u\in\mathcal{U}$$

With AGMA, the posynomial constraints **C5-C8** become monomial and $\widetilde{\mathbb{P}_2}$ is converted into a standard GP (i.e. $\widehat{\mathbb{P}_2}$), the constraints $\mathbf{C4}, \widetilde{\mathbf{C6}} - \widetilde{\mathbf{C8}}$ satisfy $\forall b\in\mathcal{B}, \forall\boldsymbol{s}\in\mathcal{S}, \forall u\in\mathcal{U}$. Finally, SCA algorithm updates the variables in each iteration l. The optimization problem under different time iterations is:

$$\widehat{\mathbb{P}_2} : \min_{\boldsymbol{W},\boldsymbol{Y},\boldsymbol{P},\boldsymbol{X}} \sum_{b\in\mathcal{B}}\sum_{u\in\mathcal{U}}\sum_{\boldsymbol{s}\in\mathcal{S}} \varphi_u \mathrm{P}_s(\boldsymbol{s}) I^{\boldsymbol{s}}_{u,b} R^{\boldsymbol{s}_\mathrm{e}}_{u,b} x^{\boldsymbol{s}}_{u,b}(l) \quad \textbf{s.t. } \widetilde{\mathbf{C1}}, \mathbf{C4}$$

$$\widetilde{\mathbf{C5}} : \sum\nolimits_{b\in\mathcal{B}}\sum\nolimits_{u\in\mathcal{U}}\sum\nolimits_{\boldsymbol{s}\in\mathcal{S}} \mathrm{P}_s(\boldsymbol{s}) I^{\boldsymbol{s}}_{u,b} R^{\boldsymbol{s}_\mathrm{r}}_{urllc,b} x^{\boldsymbol{s}}_{u,b}(l) \ge \eta$$

$$\widetilde{\mathbf{C6}} : e^{-\lambda}\Big(\frac{\alpha^{\boldsymbol{s}}_{u,b}(l)}{y^{\boldsymbol{s}}_{u,b}(l)}\Big)^{\alpha^{\boldsymbol{s}}_{u,b}(l)}\Big(\frac{\beta^{\boldsymbol{s}}_{u,b}(l)}{w^{\boldsymbol{s}}_{u,b}(l)\lambda}\Big)^{\beta^{\boldsymbol{s}}_{u,b}(l)} \le 1$$

$$\widetilde{\mathbf{C7}} : \frac{y^{\boldsymbol{s}}_{u,b}\Big(2\lambda^{-1}+1\Big)e^{-\lambda}\Big(\frac{\psi^{\boldsymbol{s}}_{u,b}(l)2\lambda}{e^{-2\lambda}(\lambda+2)}\Big)^{\psi^{\boldsymbol{s}}_{u,b}(l)}}{\Big(\frac{p^{\boldsymbol{s}}_b y^{\boldsymbol{s}}_{u,b}e^{\lambda}}{\theta^{\boldsymbol{s}}_{u,b}(l)}\Big)^{\theta^{\boldsymbol{s}}_{u,b}(l)}\Big(\frac{y^{\boldsymbol{s}}_{u,b}e^{\lambda}}{\delta^{\boldsymbol{s}}_{u,b}(l)}\Big)^{\delta^{\boldsymbol{s}}_{u,b}(l)}\Big(\frac{\big(y^{\boldsymbol{s}}_{u,b}\big)^2\big(\frac{2+\lambda}{2\lambda}\big)}{\epsilon^{\boldsymbol{s}}_{u,b}(l)}\Big)^{\epsilon^{\boldsymbol{s}}_{u,b}(l)}} \le 1$$

$$\widetilde{\mathbf{C8}} : \frac{y^{\boldsymbol{s}}_{u,b}(l)w^{\boldsymbol{s}}_{u,b}(l)^2}{\Big(\frac{y^{\boldsymbol{s}}_{u,b}(l)x^{\boldsymbol{s}}_{u,b}(l)}{\nu^{\boldsymbol{s}}_{u,b}(l)}\Big)^{\nu^{\boldsymbol{s}}_{u,b}(l)}\Big(\frac{e^{-\lambda}w^{\boldsymbol{s}}_{u}(l)^2}{\tau^{\boldsymbol{s}}_{u,b}(l)}\Big)^{\tau^{\boldsymbol{s}}_{u,b}(l)}} \le 1$$

and the iteration variables are updated as (8)-(13), and all of them satisfy $\forall b\in\mathcal{B}, \forall\boldsymbol{s}\in\mathcal{S}, \forall u\in\mathcal{U}$.

$$\alpha^{\boldsymbol{s}}_{u,b}(l) = \frac{y^{\boldsymbol{s}}_{u,b}(l)}{y^{\boldsymbol{s}}_{u,b}(l)+w^{\boldsymbol{s}}_{u,b}(l)\lambda}, \beta^{\boldsymbol{s}}_{u,b}(l) = \frac{w^{\boldsymbol{s}}_{u,b}(l)\lambda}{y^{\boldsymbol{s}}_{u,b}(l)+w^{\boldsymbol{s}}_{u,b}(l)\lambda} \tag{8}$$

$$\theta^{\boldsymbol{s}}_{u,b}(l) = p^{\boldsymbol{s}}_b y^{\boldsymbol{s}}_{u,b}e^{\lambda}\Big(\gamma^{\boldsymbol{s}}_{u,b}(l)\Big)^{-1}, \delta^{\boldsymbol{s}}_{u,b}(l) = e^{\lambda}y^{\boldsymbol{s}}_{u,b}(l)\Big(\gamma^{\boldsymbol{s}}_{u,b}(l)\Big)^{-1} \tag{9}$$

$$\epsilon^{\boldsymbol{s}}_{u,b}(l) = \frac{(y^{\boldsymbol{s}}_{u,b})^2\Big(2+\lambda\Big)}{2\lambda\gamma^{\boldsymbol{s}}_{u,b}(l)}, \psi^{\boldsymbol{s}}_{u,b}(l) = \frac{\Big(2+\lambda\Big)e^{-2\lambda}}{2\lambda\gamma^{\boldsymbol{s}}_{u,b}(l)} \tag{10}$$

where, $\gamma^{\boldsymbol{s}}_{u,b}(l) =$ (11)

$$\Big(p^{\boldsymbol{s}}_b(l)y^{\boldsymbol{s}}_{u,b}(l) + y^{\boldsymbol{s}}_{u,b}(l)\Big)e^{\lambda} + \frac{(2+\lambda)e^{-2\lambda}}{2\lambda} + (y^{\boldsymbol{s}}_{u,b}(l))^2(\frac{\lambda+2}{2\lambda})$$

$$\nu^{\boldsymbol{s}}_{u,b}(l) = y^{\boldsymbol{s}}_{u,b}(l)x^{\boldsymbol{s}}_{u,b}(l)\big(y^{\boldsymbol{s}}_{u,b}(l)x^{\boldsymbol{s}}_{u,b}(l) + e^{-\lambda}w^{\boldsymbol{s}}_{u,b}(l)^2\big)^{-1} \tag{12}$$

$$\tau^{\boldsymbol{s}}_{u,b}(l) = e^{-\lambda}w^{\boldsymbol{s}}_{u,b}(l)^2\big(y^{\boldsymbol{s}}_{u,b}(l)x^{\boldsymbol{s}}_{u}(l) + e^{-\lambda}w^{\boldsymbol{s}}_{u,b}(l)^2\big)^{-1} \tag{13}$$

C. Computational Complexity Analysis

The computational complexity of Algorithm 1 is dominated by the URLLC placement scheduling, in which we solve a series of GP sub-problems with interior point method. From [12], the order of computational complexity for $\widehat{\mathbb{P}_2}$ is $i_{urllc} \times \frac{log(C_{urllc}/(t^0\varrho))}{log(\partial)}$, where i_{urllc} is the number of required computations to locally approximate $\widehat{\mathbb{P}_2}$ with a GP using AGMA and the later represents the number of required iterations, in which t^0 is the initial point to approximate the accuracy, $0 < \varrho < 1$ is the stopping criterion and ∂ is used for

Table I: System Configuration

Parameters	Values
Transmission error rate	$\xi = 10^{-9}$
Channel block-length	$b_l = 200$
The number of channels	$B = 4$
The threshold of URLLC outage probability	$\zeta = 0.9$
The threshold of URLLC mean throughput	$\eta = 1.2$Mbps
Transmit SNR of eMBB users	$\rho = 10^4$ (i.e. 40dB)
Transmit SNR of URLLC users	$\rho_r = 10^4$ (i.e. 40dB)
Channel bandwidth	$B_W = 180$KHz

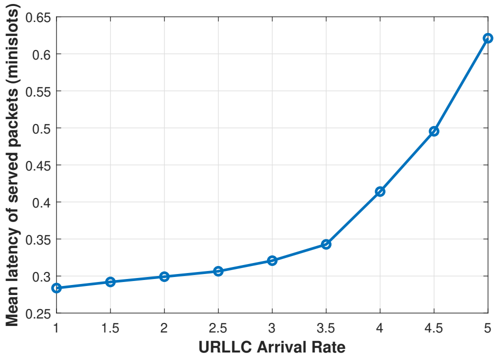

Fig. 2. Average delay of served URLLC packets vs. URLLC arrival rate.

updating the accuracy of interior point method. Also, $C_{\widehat{urllc}}$ denotes the number of URLLC scheduling constraints in $\mathbb{P}_2$ which is $C_{urllc} = 2 + 4BSU$. The number of computations for URLLC placement scheduling is $i_{urllc} = B^2SU + 5BSU$. It is worth noting that the studied scheduling problem is a stochastic problem and thus needs to be run *once*. The obtained optimal values (i.e., $w_{u,b}^{\boldsymbol{s}}$ and $I_{u,b}^{\boldsymbol{s}}$) can be applied in each time-slot and each channel to have the allocations depending on what the instantaneous channel is.

IV. Simulation Results

We consider a system with three eMBB users with the setting in Table I. We assumed Rayleigh fading model considering three different ranges of fading power gains $[0, 0.5]$, $[0.5, 1.5]$ and $[1.5, \infty]$. The PMF of possible channel states is calculated as $\int_{\vartheta_2}^{\vartheta_1} \chi g(\chi) d\chi / \int_{\vartheta_2}^{\vartheta_1} g(\chi) d\chi$, where χ, $g(\chi)$, ϑ_1 and ϑ_2 represent the value of the fading power gains, its probability density function, and its lower and upper bounds respectively. Therefore, the set of channel states $\mathcal{C} = \{0.2287, 0.9129, 2.5097\}$ and the PMF of corresponding different possible channel states is $\{0.39, 0.38, 0.23\}$.

First, we consider a case in which for all eMBB users $d = 0.3$, which is the normalized distance to the maximum radius of the cell and consider the same priority weights for all eMBB users ($\varphi_u = 1/3$). Fig. 2 numerically explores the average queuing time for served URLLC packets against URLLC arrival rate λ_a. From the results, the value of mean latency is below 0.65 mini-slots when λ_a changing within $[1, 5]$ packets/mini-slot. However, when the load rate is greater than 4 packets/mini-slot, there is a rapid rise on such delay, since the channel limitation leads to longer waiting time.

Fig. 3 depicts the trade-off between URLLC average served rate and the expected total eMBB throughput with the proposed queuing policy and resource proportional (RP) placement scheduling policy [9]. The latter policy assumes that puncturing weights are in proportion to the percentage of resource blocks that allocated to eMBB user u in each time slot. The curves have been obtained by changing λ_a and measuring eMBB throughput and URLLC served packet rate.

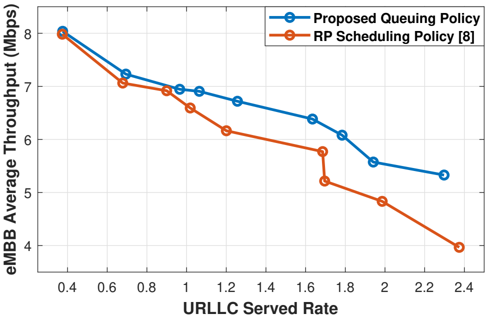

Fig. 3. Mean eMBB total throughput versus URLLC mean served rate.

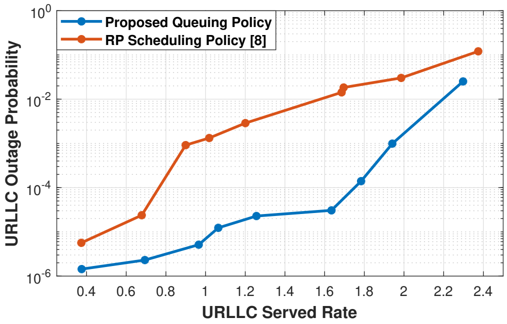

Fig. 4. URLLC outage probability vs. URLLC mean served rate.

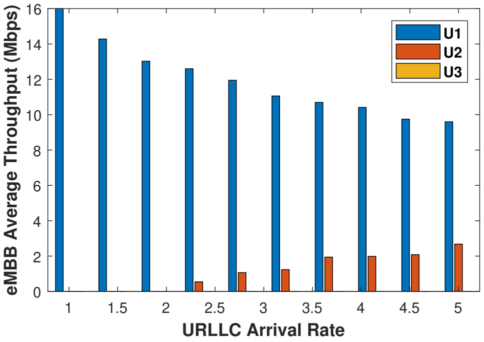

Fig. 5. eMBB throughput vs URLLC arrival rate ($\varphi_u = 1/3$, $\forall u \in \mathcal{U}$).

It is shown that with the same URLLC served rate, the system with queuing can support a higher eMBB total throughput for higher range of URLLC served rate. As λ_a being increased, the BS reduces eMBB throughput to allocate more mini slots for enhancing URLLC reliability in both systems, but the RP placement scheduling policy drops more URLLC packets. This has been also confirmed in Fig. 4 which demonstrates achievable URLLC outage probability versus URLLC served rate. Compared with the RP scheduling policy, it is clear that the queuing mechanism is effective in keeping URLLC outage probability limited even with a larger URLLC rate.

Here, we consider a setting in which eMBB users are at different distances from the BS (i.e., $d_1 = 0.1$, $d_2 = 0.5$, $d_3 = 0.9$). We change the priority weights to observe the influence on expected throughput of each eMBB user. In Fig. 5, we apply the same priority weights on eMBB users, i.e. $\varphi_u = 1/3$, $\forall u \in \mathcal{U}$. The BS allocates more time slots to the eMBB user which has the highest average channel rate (i.e., U_1) and also punctures abundant URLLC packets on it. As λ_a increases, larger puncturing weights are required to support URLLC reliability requirement and this reduces the average throughput of U_1. Thus, the BS enhances the total

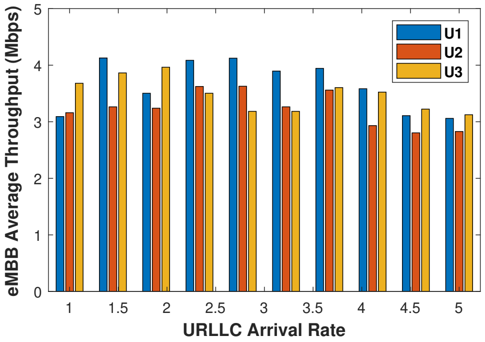

Fig. 6. eMBB throughput vs URLLC arrival rate ($\varphi_u = \{0.25, 0.35, 0.45\}$).

eMBB throughput by allocating time slots to both U_2 and U_1 under different instantaneous CSI and never allocates any time slot to the user at the edge. To improve such unfairness among eMBBs, in Fig. 6, we adjust the priority weights to achieve fair allocations. The priority weight of a user is set inversely proportional to the distance of that user from the BS, i.e., $\varphi_1 = 0.25$, $\varphi_2 = 0.35$, $\varphi_3 = 0.45$. Thus, the priority weights can adjust the allocations even if the eMBB users are at different distances from the BS.

V. Conclusion

In this paper, we developed an optimal joint scheduling strategy for the eMBB and URLLC coexistence system with a queuing mechanism. The results showed that the system mechanism satisfied URLLC reliability and latency requirements when comparing with the no queue system under RP placement policy. With suitable priority factors, the system also maintains the throughput of each eMBB user at different transmission distances.

Appendix A
Proof of Expected Value for Served URLLC Packets

We assumed all channels are the same in different channels and scheduled each channel individually. At the beginning of each mini slot, the number of URLLC served packet in each single channel is either zero or one, thus, the expected number of URLLC served packets for all transmission channels at each mini slot m when CSI is $\mathbf{s}$ can be expressed as

$$E[\mu^{\boldsymbol{s}}_m] = \sum\nolimits_{b\in\mathcal{B}} \Pr(\mu^{\boldsymbol{s}}_{m,b} = 1) \tag{14}$$

$$= \sum\nolimits_{b\in\mathcal{B}} \Pr(C^{\boldsymbol{s}}_{m,b} = 1)\Big(1 - \Pr(Q^{\boldsymbol{s}}_{m-1,b} = 0)\Pr(A^{\boldsymbol{s}}_{m,b} = 0)\Big)$$

We assume the URLLC packets as Possion distribution, therefore, $\Pr(A^{\boldsymbol{s}}_{m,b} = 0) = e^{-\lambda}$. The problem becomes to find the distribution of the provided mini slots $\big($i.e., $\Pr(C^{\boldsymbol{s}}_{m,b} = c)$, $\forall c \in \{0,1\}\big)$ and the distribution of queue $\big($i.e., $\Pr(Q^{\boldsymbol{s}}_{m,b} = q)$, $\forall q \in \{0, 1, \ldots, \infty\}\big)$ for each channel b.

The PMF of provided mini slots for channel b, $\forall b \in \mathcal{B}$ is

$$\Pr(C^{\boldsymbol{s}}_{m,b} = c) = (w^{\boldsymbol{s}}_b)^c\big(1 - (w^{\boldsymbol{s}}_b)\big)^{1-c}, \forall c \in \{0,1\} \tag{15}$$

where $w^{\boldsymbol{s}}_b$ the total puncturing weights for channel b when CSI is $\mathbf{s}$ and can be calculated with

$$w^{\boldsymbol{s}}_b = \mathrm{diag}(\boldsymbol{W_b}^T\boldsymbol{I_b}) = \sum_{u\in\mathcal{U}} w^{\boldsymbol{s}}_{u,b} I^{\boldsymbol{s}}_{u,b}, \forall \boldsymbol{s} \in \mathcal{S}, \forall b \in \mathcal{B} \tag{16}$$

The state transition matrix is used to find the steady state probability of the queue system. The equation for the state transition matrix for each sub queue is in Fig.7. The matrix obeys the steady state equation as

$$\boldsymbol{p}^T = \begin{pmatrix} \Pr(A_{m,b}{=}0) + \Pr(A_{m,b}{=}1)\Pr(C^s_{m,b}{=}1) & \Pr(A_{m,b}{=}0) & \Pr(A_{m,b}{=}0) & \\ \Pr(A_{m,b}{=}1)\Pr(C^s_{m,b}{=}0) + \Pr(A_{m,b}{=}2)\Pr(C^s_{m,b}{=}1) & \Pr(A_{m,b}{=}1) & \Pr(A_{m,b}{=}1) & \cdots \\ \vdots & \vdots & \vdots & \cdots \\ \Pr(A_{m,b}{=}k)\Pr(C^s_{m,b}{=}0) + \Pr(A_{m,b}{=}k{+}1)\Pr(C^s_{m,b}{=}1) & \Pr(A_{m,b}{=}k) & \Pr(A_{m,b}{=}k) & \cdots \\ \vdots & \vdots & \vdots & \\ \Pr(A_{m,b}{=}\infty)\Pr(C^s_{m,b}{=}0) & \Pr(A_{m,b}{=}\infty) & \Pr(A_{m,b}{=}\infty) & \end{pmatrix}$$

Fig. 7. The state transition matrix for channel b.

$$(\mathrm{I} - \mathbf{P}^T)[\Pr(Q^{\boldsymbol{s}}_{m,b} = 0), \ldots, \Pr(Q^{\boldsymbol{s}}_{m,b} = \infty)]^T = \mathbf{0} \tag{17}$$

The sum of all steady state probabilities equals to one, i.e.,

$$\sum\nolimits_{q\in\{0,\ldots,\infty\}} \Pr(Q^{\boldsymbol{s}}_{m,b} = q) = 1 \tag{18}$$

Therefore, the steady state probabilities for $\Pr(Q^{\boldsymbol{s}}_{m,b} = 0)$ at channel b is

$$\Pr(Q^{\boldsymbol{s}}_{m,b} = 0) = \sum_{u\in\mathcal{U}} I^{\boldsymbol{s}}_{u,b}(e^{-\lambda} - \lambda w^{\boldsymbol{s}}_{u,b})^{-1}, \forall m \in \mathcal{M} \tag{19}$$

With (14)-(19), $E[\mu^s_m]$ can be derived as (5).

Appendix B
Proof of Probability for No Dropped URLLC Case

The formula of dropped packets on channel b in (3) can be written as

$$D^{\boldsymbol{s}}_{m,b} = \begin{cases} 0, & \mu^{\boldsymbol{s}}_{m,b} = T^{\boldsymbol{s}}_{m,b} \\ \max(Q^{\boldsymbol{s}}_{m-1,b} - \mu^{\boldsymbol{s}}_{m,b}, 0), & \mu^{\boldsymbol{s}}_{m,b} = C^{\boldsymbol{s}}_{m,b} \end{cases}$$

Thus, based on (15)-(19), $\Pr(D^{\boldsymbol{s}}_{m,b} = 0)$ can be derived as $\Pr(D^{\boldsymbol{s}}_{m,b} = 0) = \Pr(Q^{\boldsymbol{s}}_{m-1,b} = 0) + \Pr(Q^{\boldsymbol{s}}_{m-1,b} = 1)\Pr(C^{\boldsymbol{s}}_{m} = 1)$

$$= \sum_{u\in\mathcal{U}} I^{\boldsymbol{s}}_{u,b} \frac{2\lambda w^{\boldsymbol{s}}_{u,b} - \big(w^s_{u,b}\big)^2 e^{-\lambda}\big(2\lambda + \lambda^2\big)}{2e^{-\lambda} - w^{\boldsymbol{s}}_{u,b}\lambda}, \forall \boldsymbol{s} \in \mathcal{S}, \forall b \in \mathcal{B}$$

References

[1] Y. Zikria et al, "5G mobile services and scenarios: Challenges and solutions," *Sustainability*, vol. 10, p. 3626, 2018.

[2] 3GPP, "New services and applications with 5G ultra-reliable low latency communication," *5G Americas WHITEPAPER*, November 2018.

[3] M. Luvisotto et al, "Ultra high performance wireless control for critical applications: Challenges and directions," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 3, pp. 1448–1459, 2017.

[4] 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 Wireless Communications*, vol. 25, no. 3, pp. 124–130, 2018.

[5] M. Bennis, M. Debbah, and H. V. Poor, "Ultra-reliable and low-latency wireless communication: Tail risk and scale," *Proc. IEEE*, vol. 106, no. 10, pp. 1834–1853, Oct. 2018.

[6] C. Li et al, "5G ultra-reliable and low-latency systems design," in *2017 EuCNC*, 2017, pp. 1–5.

[7] K. I. Pedersen et al, "Punctured scheduling for critical low latency data on a shared channel with mobile broadband," in *IEEE VTC-Fall*, 2017.

[8] 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, 2020.

[9] T. R. Pijnappel, S. C. Borst, and P. A. Whiting, "Joint scheduling of low-latency and best-effort flows in 5G wireless networks," in *2020 18th WiOpt*, 2020, pp. 1–8.

[10] M. Alsenwi et al, "embb-urllc resource slicing: a risk-sensitive approach," *IEEE Communications Letters*, vol. 23, no. 4, pp. 740–743, 2019.

[11] Y. Huang, S. Li, C. Li, Y. T. Hou, and W. Lou, "A deep-reinforcement-learning-based approach to dynamic embb/urllc multiplexing in 5G NR," *IEEE Internet of Things Journal*, vol. 7, no. 7, pp. 6439–6456, 2020.

[12] S. Parsaeefard, R. Dawadi, M. Derakhshani, and T. Le-Ngoc, "Joint user-association and resource-allocation in virtualized wireless networks," *IEEE Access*, vol. 4, pp. 2738–2750, 2016.

[13] G. Durisi, T. Koch, and P. Popovski, "Toward massive, ultra reliable, and low-latency wireless communication with short packets," *Proceedings of the IEEE*, vol. 104, no. 9, pp. 1711–1726, 2016.