Joint Resource and Power Allocation for URLLC-eMBB Traffics Multiplexing in 6G Wireless Networks

Mohammed Almekhlafi,[1] Mohamed Amine Arfaoui,[1] Chadi Assi[1] and Ali Ghrayeb[2]
[1]Concordia University, Montreal, Canada, emails: { m_almekh@encs, m_arfaou@encs, assi@ciise}.concordia.ca
[2]Texas A&M University at Qatar, Doha, Qatar, email: ali.ghrayeb@qatar.tamu.edu

ICC 2021 - IEEE International Conference on Communications | 978-1-7281-7122-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICC42927.2021.9500443

***Abstract*—Ultra-Reliable and Low Latency Communications (URLLC) is one of the essential services in 5G networks and beyond. The coexistence of URLLC alongside other service classes, namely, enhanced Mobile BroadBand (eMBB) and massive Machine-Type Communications (mMTC), calls for developing spectrally efficient multiplexing techniques. In this work, we study the problem of scheduling URLLC traffic in a downlink system with the presence of eMBB traffic class. Based on the superposition/puncturing scheme, a resource allocation problem is formulated with the objective to minimize the eMBB data rate loss while satisfying eMBB and URLLC quality of service (QoS) constraints. The resulting problem is formulated as a mixed integer non-linear programming (MINLP) which is generally NP hard and hence complex to solve. Hence, we derive its feasibility region as well as the optimal solutions for the power and spectral resource allocation. Subsequently, we propose a low complexity algorithm to serve URLLC traffic. Simulation results show that the proposed algorithm achieves higher reliability for URLLC and higher eMBB data rate compared to the puncturing schemes. The results also show that the eMBB QoS requirements, which are represented by the eMBB rate loss threshold, has a negative effect on the URLLC reliability for high URLLC load. Therefore, the eMBB rate and the eMBB loss threshold should be jointly optimized considering QoS of both eMBB and URLLC.**

***Index Terms*—eMBB, multiplexing, puncturing, superposition, URLLC, 6G.**

I. Introduction

Future generations of communication networks, especially the sixth generation (6G), will inherent various services from their fifth generation (5G) predecessor, such as ultra-reliable and low-latency communications (URLLC), enhanced Mobile BroadBand (eMBB) and massive Machine-Type Communications (mMTC) [1]. Due to this, these networks require to support the resulting limitations of the aforementioned services. Specifically, gigabits per second (Gbps) data rates and millisecond latency will soon become insufficient, given the exponential growth of mobile data traffic and the rapid proliferation of time-critical services [2]–[4]. Nevertheless, the ambitious vision of 6G aims toward enabling services with terabits per second data rates and sub-millisecond latency. Therefore, 6G is expected to support some URLLC applications that focus on simultaneously supporting massive connections and high data rates instead of sparse and short packet transmissions [4]. As a result, these high requirements complicate the co-existence of emerging URLLC services with their eMBB and mMTC counterparts.

Enabling support for URLLC services has received considerable research interests in the past recent years [1], [5], [6]. Traffic belonging to the URLLC service class requires immediate scheduling and transmission once it arrives to the base Station (BS), which translates into immediate availability of spectral resources. Hence, on-air resource allocation for URLLC data has been shown to be more spectrally effective than reservation-based scheduling [5], [6]. In line with the on-air allocation, the superposition/puncturing schemes have been proposed by the Third Generation Partnership Project (3GPP) standard [6]. Arriving URLLC packets are immediately transmitted in the next mini-slot (0.125 ms) over the ongoing eMBB resources once arrived at the transmitting BS. In this context, the puncturing technique consists of puncturing a part of the frequency resources, which are already allocated to an eMBB user at the beginning of each time-slot, and re-allocating this resource part to some URLLC packets. The main drawback of the puncturing scheme is that the punctured resources can significantly reduce the throughput of eMBB service classes due to high overhead and retransmissions [5], [7]. On the other hand, the superposition technique consists of superimposing the URLLC and eMBB packets within a part of the frequency resources, which are already allocated to an eMBB user at the beginning of each time-slot. This technique can be performed using superposition coding (SC), which is a main component of non-orthogonal multiple access (NOMA). The main drawback of the superposition scheme is that it impacts the reliability of URLLC packets due to the interference caused by the eMBB traffic.

Based on the superposition/puncturing scheme, several approaches focusing on scheduling URLLC traffic with the aim of maximizing the total average data rate of eMBB users have been proposed [6]–[10]. The authors of [6] studied the joint eMBB and URLLC scheduling problem, and they considered linear, convex and threshold models for the eMBB rate loss resulting from the superposition/puncturing scheme. A resource allocation policy for a puncturing-based scheduler was proposed in [7], where the formulated problem considered the overhead associated with the URLLC load segmentation while maximizing the rate utility. In [8], a risk-sensitive approach was introduced to alleviate the puncturing effects on the eMBB users with low data rates. In [9], a deep reinforcement learning approach was proposed to allocate the URLLC traffic. The work in [7]–[9] considered only the puncturing scheme for joint scheduling of eMBB and URLLC loads. Authors in [10] have formulated a URLLC traffic allocation problem by adopting a superposition or puncturing scheme. Against

the above background, there is a lack of considering the eMBB quality-of-service (QoS), i.e., eMBB rate requirements, for both the puncturing and the superposition schemes. Particularly, eMBB QoS requirements can effect the URLLC reliability, and therefore it should be carefully investigated.

In this work, we investigate the performance of a superposition/puncturing allocation scheme in a downlink system that consists of a single base station serving multiple eMBB and URLLC users. The objective of the allocation problem is to minimize the eMBB rate loss due to the superposition/puncturing scheme while satisfying the eMBB and URLLC QoS requirements. Accordingly, we formulate this objective as an optimization problem, namely a mixed-integer nonlinear program (MINLP), that is generally hard to be solved. We then reformulate the problem as a bi-level optimization problem which consists of one inner problem which aims to find the optimal power and frequency resources for each URLLC and eMBB pair, and one outer problem which aims to find the optimal eMBB-URLLC pairing policy. In the inner problem, we derive the feasibility conditions and the optimal frequency and power allocation scheme in closed-form expressions. The outer problem is reduced to a simple assignment problem that can be optimally solved by using a greedy algorithm which has polynomial time complexity. The performance of the proposed solution is compared with two puncturing baselines proposed in the literature [6], [11]. Our simulation results show that the proposed algorithm achieves better eMBB loss and URLLC reliability compared to the puncturing baselines. Simulation results also show that the eMBB QoS requirements, i.e., eMBB rate loss threshold, has a bad affect on the URLLC reliability while increasing URLLC load and URLLC packet size. Hence, the eMBB rate and the eMBB loss threshold should be jointly optimized considering QoS of both eMBB and URLLC services.

The rest of the paper is organized as follows. Section II presents the system model. Section III presents the problem formulation. Section IV presents the proposed solution approach. Sections V and VI presents the simulation results and the conclusion, respectively.

II. System Model

We consider a downlink radio access network (RAN) which consists of a single base station (BS) with B resource blocks (RBs) of bandwidth W. The BS serves sets of eMBB and URLLC traffic denoted by $\mathcal{E} = \{1, 2, \ldots, E\}$ and $\mathcal{U} = \{1, 2, \ldots, U\}$, respectively (see Fig. 1). The channel gains of the eMBB users and URLLC users are denoted by $H = \{h_1, h_2, \ldots, h_E\}$ and $G = \{g_1, g_2, \ldots, g_U\}$, respectively. Time is divided into slots, and each time-slot is further divided into, $\mathcal{N} = \{1, 2, \ldots, N\}$, mini-slots of duration δ to support the latency requirement of the URLLC traffic. We assume that the URLLC user request follows a Bernoulli distribution with probability p, with fixed packet size, i.e., information bits ζ. Then, the URLLC load follows a Binomial distribution with mean $|U|p$, where $|U|$ is the cardinality of the set of URLLC users. The available resources B are allocated to the

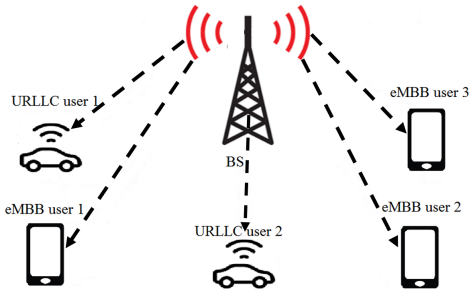


Fig. 1: System Model

eMBB traffic at the beginning of each slot, while the arriving URLLC traffic immediately superposes/punctures the eMBB resources in the next mini-slot $n \in \mathcal{N}$. We also assume that the eMBB traffic is allocated using orthogonal multiple access (OMA) in slot basis while the URLLC load is allocated based on the superposition/puncturing scheme in mini-slot basis. The superposed eMBB resources is allocated based on power domain. In superposition, the power is divided between the users sharing the same resources. In order to achieve better URLLC reliability, it is assumed that the BS allocates more power to the URLLC user. For more clarity, allocating more power to the URLLC traffic guarantees higher rates for users in this service class and less complexity at the receiver. The less complexity is due to the fact that the URLLC receiver does not perform successive interference cancellation (SIC), which may lead to violate the URLLC reliability.

A. Signal and wireless Model

We assume the BS assigns its resources at the beginning of each time slot to the eMBB users using orthogonal resources. Accordingly, the transmitted signal for eMBB user e is:

$$x = \sqrt{P} x_e, \tag{1}$$

where P is the average transmitted power and x_e is the eMBB signal. The eMBB achievable rate is then:

$$r_e = \log_2 (1 + \gamma_e) \tag{2}$$

where $\gamma_e = \frac{P|h_e|^2}{\sigma^2}$ and σ is the AWGN noise level of the eMBB receiver. In the scenarios of URLLC and eMBB coexistence, the eMBB signal is superimposed/punctured by the URLLC signal. Let l^n be the transmitted URLLC packets at mini-slot n. Then, the superimposed signal (of both eMBB user e and URLLC packet $l \in l^n$) at time mini-slot n is given as [12]:

$$x^n = \sqrt{P}(\sqrt{\alpha^n_{e,l}} x^n_l + \sqrt{1 - \alpha^n_{e,l}} x_e), \tag{3}$$

where x^n_l and $\alpha^n_{e,l}$ are the transmitted URLLC signal and the power allocation factor of user l, respectively. To accommodate the URLLC reliability requirements, the URLLC power allocation factor is assumed to be $0.5 < \alpha^n_{e,l} \leq 1$. Accordingly, the eMBB user uses SIC to cancel the signal of the URLLC and decode its signal free of interference at a rate of [12]:

$$r^n_{e,l}(\alpha^n_{e,l}) = \log_2 \left(1 + (1 - \alpha^n_{e,l})\gamma_e\right), \tag{4}$$

On the other hand, the URLLC user can decode its signal without performing SIC by treating the eMBB signal as interference [12]. As URLLC packet is small, its achievable rate follows the finite block length regime [13]. Hence, the URLLC achievable rate, for a URLLC packet superposes $\varphi^n_{e,l}$ eMBB frequency resources, is:

$$C^n_{e,l} = \log_2\left(1+\frac{\alpha^n_{e,l}\gamma^n_l}{(1-\alpha^n_{e,l})\gamma^n_l+1}\right) - \sqrt{\frac{V}{\delta\varphi^n_{e,l}W}}\frac{Q^{-1}(\epsilon_u)}{\ln(2)} \quad (5)$$

where $\gamma^n_l = \frac{P|g^n_l|^2}{\sigma^2}$ and σ is the the AWGN noise level of the URLLC receiver. Also, ϵ_u and V are the URLLC block error rate and the channel dispersion. Where $V = 1 - \frac{1}{(1+\frac{\alpha^n_{e,l}\gamma^n_l}{(1-\alpha^n_{e,l})\gamma^n_l+1})^2}$, and $V = 1 - \frac{1}{1+\frac{\alpha^n_{e,l}}{1-\alpha^n_{e,l}}} \approx 1$ at high SNR and high $\alpha^n_{e,l}$, which is the case of URLLC [14]. Noting that puncturing is assumed as a special case of superposition when the URLLC power allocation factor $\alpha^n_{e,l} = 1$.

III. Problem Formulation

Our objective is to minimize the eMBB rate loss while satisfying QoS requirements of the eMBB and the URLLC users. Let ϕ_e and φ^n_l be the RBs allocated to the to eMBB user e, and the URLLC packet l at mini-time slot n, respectively. Also, consider that $\varphi^n_l = \sum_e \varphi^n_{e,l}$ determines the RBs allocated to the URLLC packet l at mini-slot n, where $\varphi^n_{e,l}$ is the superposed/punctured resources of the eMBB user e by URLLC packet l. Then, the achievable rate, of the superimposed/punctured part of eMBB user by URLLC packet l user at mini-slot n, is:

$$R^n_{e,l} = \frac{W}{N}\varphi^n_{e,l}r^n_{e,l} \quad (6)$$

Accordingly, the rate loss for eMBB user at mini-slot n can be expressed as:

$$\widehat{R}^n_{e,l} = \frac{W}{N}\left(r_e - r^n_{e,l}\left(\alpha^n_{e,l}\right)\right)\varphi^n_{e,l} \quad (7)$$

where $\widehat{(.)}$ represents the rate loss. Therefore, the objective of minimizing the eMBB rate loss is formulated as:

$$\min_{\varphi,\alpha} \sum_e^{|E|}\sum_l^{l^n} \widehat{R}^n_{e,l} \quad (8)$$

A. eMBB and URLLC QoS

The superposition/puncturing of eMBB resources impacts the rate of the eMBB user and its reliability. Hence, we consider a threshold loss function such that the eMBB code-word will be decoded correctly with certain block error rate if the eMBB rate loss is below R^{th}_e. The loss function is derived as follows:

$$\hat{R}^n_e = \sum_l^{l^n} \widehat{R}^n_{e,l} \leq \widehat{R}^{th}_e \quad (9)$$

where $\widehat{R}^n_l$ depends on the rate of the eMBB user. Equation (9) indicates that if the eMBB loss exceeds R^{th}_e, the eMBB code-word will be received in error with a high probability, i.e., a retransmission is required.

The URLLC traffic requires high reliability and is subject to latency constraints which should be satisfied. Actually, several components, including queueing delay and transmission time, can impact the URLLC latency. The queueing delay is eliminated by transmitting the URLLC packet immediately upon arrivals, i.e., in the next mini-slot. While the transmission delay can be guaranteed by controlling the transmission rate that satisfies the reliability requirement. Accordingly, the minimum rate that satisfies the target reliability can be expressed as:

$$C^n_{e,l} \geq C^{th}, \quad \forall n,l \qquad (bits/sec/Hz) \quad (10)$$

where C^{th} is the minimum URLLC rate. Also, a URLLC packet should be transmitted within one mini-slot according to the constraint in (10):

$$W\delta\varphi^n_l C^n_{e,l} \geq \zeta \quad (11)$$

Accordingly, the URLLC packet has a reliable transmission if (10) and (11) are satisfied. Therefore, the final optimization problem of the URLLC schedulers is formulated as follows, at each mini-slot belonging to slot $n >= 2$:

$$\mathcal{P}: \min_{\mathbf{I},\boldsymbol{\varphi},\boldsymbol{\alpha}} \sum_e^{|E|}\sum_l^{l^n} I^n_{e,l}\widehat{R}^n_{e,l} \quad (12a)$$

$$\text{s.t: } \sum_e^{|E|} I^n_{e,l}C^n_{e,l} \geq C^{th}, \quad \forall l \in l^n \quad (12b)$$

$$\sum_e^{|E|} I^n_{e,l}W\delta\varphi^n_{e,l}C^n_{e,l} \geq \zeta, \quad \forall l \in l^n, \forall \quad (12c)$$

$$\sum_l^{l^n} I^n_{e,l}\widehat{R}^n_{e,l} \leq \widehat{R}^{th}_e - \sum_i^{n-1}\widehat{R}^i_{e,l}, \quad \forall e \in \mathcal{E} \quad (12d)$$

$$0.5 < \alpha^n_{e,l} \leq 1, \quad \forall l \in l^n \quad (12e)$$

$$0 \leq \varphi^n_{e,l} \leq \phi_e, \quad \forall e \in \mathcal{E} \quad (12f)$$

$$\sum_e^{|E|} I^n_{e,l} = 1, \forall l \in l^n \quad (12g)$$

$$I^n_{e,l} \in \{0,1\} \forall e \in \mathcal{E}, \forall l \in l^n \quad (12h)$$

where $\mathbf{I}$ is a matrix of dimensions $E \times U$ which represents the pairing decision variables. The above problem seeks the optimal pairing of URLLC packet and eMBB users and the optimum resource allocation vector $\boldsymbol{\varphi}$ and the power allocation vector $\boldsymbol{\alpha}$ to minimize the sum of eMBB rate loss (12a). Constraints (12b) and (12c) ensure the URLLC packets reliability. Constraint (12d) represents the QoS of the eMBB users. The bounds for the decision variables are ensured in (12e), (12f), (12g) and (12h). However, the problem $\mathcal{P}$ is a mixed integer non-linear problem (MINLP), which is generally very hard to solve, understanding the relation between the decision variables $I^n_{e,l}$,$\varphi^n_{e,l}$ and $\alpha^n_{e,l}$ can further help to simplify the optimization problem $\mathcal{P}$.

IV. Solution Approach

Following the above discussion, problem $\mathcal{P}$ has a nice property which can be exploited to solve the problem efficiently. Precisely, the resource allocation policy is independent from the pairing variable $I^n_{e,l}$. Particularly, let us consider that $\{I^n_{e,l}{}^*, \alpha^n_{e,l}{}^*, \varphi^n_{e,l}{}^* | e \in \mathcal{E}, l \in l^n\}$ represents the set of optimal user pairing and resource allocation policies of problem $\mathcal{P}$. If $I^n_{e,l}{}^* = 1$ then the optimal resource allocation $\alpha^n_{e,l}{}^*, \varphi^n_{e,l}{}^*$ can be obtained for each pair (e,l) independent of the pairing policy $I^n_{e,l}$. In other words, assuming that all (e,l) can be paired together, the optimal resource allocation $(\alpha^n_{e,l}{}^*, \varphi^n_{e,l}{}^*)$ can be obtained. Then the pairing problem $\mathcal{P}$ becomes a linear

assignment problem which determines the optimal pairing policy ${I_{e,l}^n}^*$. Accordingly, we decompose the problem $\mathcal{P}$ into two sub-problems; a resource allocation problem that minimizes the eMBB rate loss within eMBB-URLLC pairs; and an assignment problem that minimizes the total eMBB rate loss.

A. Resource allocation problem

In this section, a resource allocation problem is formulated whose objective is to find the optimal couple $\left({\alpha_{e,l}^n}^*, {\varphi_{e,l}^n}^*\right)$ for each eMBB-URLLC (e,l) which minimizes the eMBB rate loss. This problem is formulated as follows:

$$\mathcal{P}_1^{inner}: \min_{\varphi_{e,l}^n, \alpha_{e,l}^n} \widehat{R}_{e,l}^n \tag{13a}$$

$$\text{s.t: } C_{e,l}^n \geq C^{th}, \quad \forall l \in l^n, \forall \tag{13b}$$

$$W\delta\varphi_{e,l}^n C_{e,l}^n \geq \zeta, \quad \forall l \in l^n, \forall \tag{13c}$$

$$\widehat{R}_{e,l}^n \leq \widehat{R}_e^{th} - \textstyle\sum_{i=1}^{n-1} \widehat{R}_{e,l}^i, \quad \forall e \in \mathcal{E} \tag{13d}$$

$$0.5 < \alpha_{e,l}^n \leq 1, \quad \forall l \in l^n \tag{13e}$$

$$0 \leq \varphi_{e,l}^n \leq \phi_e, \quad \forall e \in \mathcal{E} \tag{13f}$$

We first investigate the feasibility conditions of $\mathcal{P}_1^{inner}$ which are defined by the following theorem:

***Theorem* 1.** *The problem $\mathcal{P}_1^{inner}$ is said to be feasible if and only if the following conditions hold:*

$${\alpha_{e,l}^n}^{min} \leq 1 \tag{14}$$

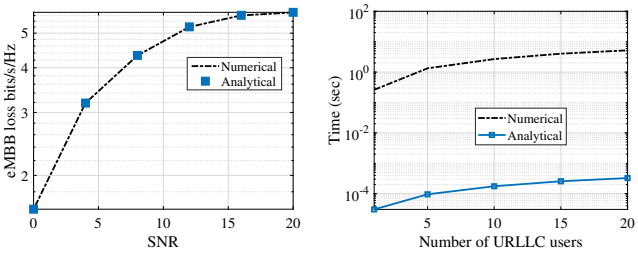
$$\left\lceil \max(0, F_{e,l}^n(1), \mathcal{F}_{e,l}^n(1)) \right\rceil \leq \min(\phi_e, \beta_{e,l}^n|_{\alpha_{e,l}^n=0.5}) \tag{15}$$

where ${\alpha_{e,l}^n}^{min}$, $F_{e,l}^n$, $\mathcal{F}_{e,l}^n$ and $\beta_{e,l}^n$ are defined in the top of the next page.

Proof. The proof can easily be derived by observing the bounds of the variable $\alpha_{e,l}^n$ and $\varphi_{e,l}^n$. Let us first investigate constraints (13d), and (13f) which together define the upper bound for $\varphi_{e,l}^n$. Constraints (13d) can be equivalently transformed into the following inequality:

$$\beta_{e,l}^n(\alpha_{e,l}^n) \leq \frac{N}{W} \frac{\widehat{R}_e^{th} - \sum_{i=1}^{n-1} \widehat{R}_{e,l}^i}{r_e - r_e(\alpha_{e,l}^n)} \tag{16}$$

By substituting $\alpha_{e,l}^n = 0.5$, we obtain the upper bound of $\varphi_{e,l}^n$ which is also bounded by (13f). After defining the maximum frequency resources that can be superposed/punctured. We can get the minimum feasible $\alpha_{e,l}^n$ that satisfy both URLLC rate constraints on (13b), and (13c) by substituting ${\varphi_{e,l}^n}^{\max} = \min(\phi_e, \beta_{e,l}^n|_{\alpha_{e,l}^n=0.5})$. By manipulating the expression, one can derive the feasibility on this variable (14). Moreover, constraints (13b) and (13c) define together lower boundaries of the region of the feasible $\varphi_{e,l}^n$ and which should be between $[0, \phi_e]$. If we substitute $y = \sqrt{\varphi_{e,l}^n}$ in (13b) and (13c), we obtain $\varphi_{e,l}^n \geq \mathcal{F}_{e,l}^n$ and $\varphi_{e,l}^n \geq F_{e,l}^n$, respectively. Solving (18) and (19) at the intersection point with $\alpha_{e,l}^n = 1$, we got the lower bound of feasible $\varphi_{e,l}^n$. In order for the inner problem to be feasible, the feasibility region should be non-empty, i.e., $\varphi_{e,l}^n$ should be greater then $\max(\mathcal{F}_{e,l}^n, F_{e,l}^n)$, and it should be less than $\min(\beta_{e,l}^n, \phi_e)$. This completes the proof. □



(a) Analytical and numerical eMBB loss rate versus SNR.

(b) Analytical and numerical time complexity.

Fig. 2: Analytical and numerical performance analysis.

Now, assuming that $\mathcal{P}_1^{inner}$ is feasible, the optimal resource allocation policy is presented in the following theorem.

***Theorem* 2.** *For the case of one-to-one pairing, the optimal resource allocation policy ${\varphi_{e,l}^n}^*$, that maximize $\mathcal{P}_1^{inner}$ is.*

$${\varphi_{e,l}^n}^* = \arg\max_i \alpha_i^* < 1 - \frac{1+\sqrt{(1+\gamma_e)((1-\alpha_{i-1}^*)\gamma_e+1)}}{\gamma_e} \tag{20}$$

Proof. After determining the feasibility condition of $\mathcal{P}_1^{inner}$, , we can proceed to find the optimal ${\varphi_{e,l}^n}^*$ and ${\alpha_{e,l}^n}^*$. Let us first consider the eMBB loss at φ_1 and $\varphi_1 + 1$:

$$\widehat{R}_e^1 = W\frac{\varphi_1}{N}\log_2\left(\frac{1+\gamma_e}{(1-\alpha_1)\gamma_e+1}\right) \tag{21}$$

$$\widehat{R}_e^2 = W\frac{\varphi_1+1}{N}\log_2\left(\frac{1+\gamma_e}{(1-\alpha_2)\gamma_e+1}\right) \tag{22}$$

Then, ${\varphi_{e,l}^n}^*$ is said to be optimal if and only if $\widehat{R}_e^1 \geq \widehat{R}_e^2$. Solving this inequality, one can derive:

$$\alpha_2 < 1 + \frac{1-\sqrt{(1+\gamma_e)((1-\alpha_1)\gamma_e+1)}}{\gamma_e} \tag{23}$$

From the upper expression, the optimal ${\varphi_{e,l}^n}^*$ is the maximum $\varphi_{e,l}^n$ that satisfy (23). This completes the proof. □

For one eMBB-URLLC pair, Fig. 2a illustrates the analytical average eMBB loss obtained through the closed-form expression derived in Theorem 2 and numerical results versus the SNR. The numerical results are obtained by solving problem $\mathcal{P}_1^{inner}$ using an off-the-shelf optimization solver[1]. This figure shows that the analytical results match perfectly the numerical results, which validate the optimality of the expression derived in Theorem 2. Fig. 2b presents the computation time for the closed form expression and the numerical solution against the number of URLLC users. The figure shows that the proposed algorithm has a processing time in the order of sub-millisecond which is suitable to the constraints of the URLLC traffic. However, obtaining numerical results will violate the URLLC latency and reliability requirements as the processing time is in the order of seconds. This renders numerical methods for obtaining solutions impractical.

[1]The used solver is generic algorithm which is a predefined matlab solver [15].

$$\alpha_{e,l}^{n\ min} = \max\left(0.5,\ \frac{\gamma_1^{th}\left(\gamma_{e,l}^n+1\right)}{\gamma_{e,l}^n\left(\gamma_1^{th}+1\right)},\ \frac{\gamma_2^{th}\left(\gamma_{e,l}^n+1\right)}{\gamma_{e,l}^n\left(\gamma_2^{th}+1\right)}\right) \tag{17}$$

$$\mathcal{F}_{e,l}^n(\alpha_{e,l}^n) = \left(\frac{Q^{-1}(\epsilon_u)\sqrt{V}}{\ln(2)\sqrt{\delta W}(\log_2(1+\frac{\alpha_{e,l}^n\gamma_l^n}{(1-\alpha_{e,l}^n)\gamma_l^n+1}) - C^{th})}\right)^2 \tag{18}$$

$$F_{e,l}^n(\alpha_{e,l}^n) = \frac{1}{4}\left(\frac{Q^{-1}(\epsilon_u)}{\ln(2)}\sqrt{\frac{V}{\delta W}}\frac{1}{\log_2(1+\frac{\alpha_{e,l}^n\gamma_l^n}{(1-\alpha_{e,l}^n)\gamma_l^n+1})} + \sqrt{(\frac{Q^{-1}(\epsilon_u)}{\ln(2)})^2\frac{V}{\delta W}\frac{1}{\log_2(1+\frac{\alpha_{e,l}^n\gamma_l^n}{(1-\alpha_{e,l}^n)\gamma_l^n+1})^2} + \frac{4\zeta}{\delta W\log_2(1+\frac{\alpha_{e,l}^n\gamma_l^n}{(1-\alpha_{e,l}^n)\gamma_l^n+1})}}\right)^2 \tag{19}$$

where $\gamma_1^{th} = 2^{C^{th}+\frac{Q^{-1}}{log(2)}\sqrt{\frac{V}{W\delta\varphi_{e,l}^{n\max}}}} - 1$ and $\gamma_2^{th} = 2^{\frac{\zeta}{W\delta\varphi_{e,l}^{n\max}}+\frac{Q^{-1}}{log(2)}\sqrt{\frac{V}{W\delta\varphi_{e,l}^{n\max}}}} - 1$

B. Optimal pairing problem

The purpose of this section is to find the optimal pairing policy $I_{e,l}^{n\,*}$. Using results of optimal resource allocations $\left(\alpha_{e,l}^{n\,*}, \varphi_{e,l}^{n\,*}\right)$ for each URLLC-eMBB pairs, the outer (resource assignment) problem aims to minimize the total eMBB rate loss is formulated as follows:

$$\mathcal{P}_1^{outer}: \min_{\mathbf{I}} \sum_e^{|E|}\sum_l^{l^n} I_{e,l}^n \times \widehat{R}_{e,l}^n\left(\alpha_{e,l}^{n\,*}, \varphi_{e,l}^{n\,*}\right) \tag{24a}$$

$$\text{s.t:} \sum_l^{l^n} I_{e,l}^n \widehat{R}_{e,l}^n\left(\alpha_{e,l}^{n\,*}, \varphi_{e,l}^{n\,*}\right) \le \widehat{R}_e^{th} - \sum_i^{n-1}\widehat{R}_{e,l}^i, \forall e\in\mathcal{E} \tag{24b}$$

$$\sum_e^{|E|} I_{e,l}^n = 1, \quad \forall l \in l^n \tag{24c}$$

$$I_{e,l}^n \in \{0,1\}, \forall e\in\mathcal{E}, \quad \forall l\in l^n, \tag{24d}$$

C. Proposed algorithm

The $\mathcal{P}_1^{outer}$ is an assignment problem which can be easily solved. To satisfy URLLC latency, we propose a low complexity algorithm to allocate/pair the URLLC users. The algorithm uses two observations that help to allocate the URLLC load. The eMBB loss is bounded by R_e^{th}, hence starting the allocation over the weak eMBB user leads to the minimum loss, especially for low URLLC load. Moreover, starting the allocation by the strong URLLC users increases the RLLC reliability as it needs less resources than the weaker URLLC users which increases the probability to allocate more URLLC packets. Accordingly, Algorithm 1 starts by sorting the eMBB and the URLLC packets such that $\gamma_1 \le \gamma_2 \cdots \le \gamma_E$ and $\gamma_1^n \ge \gamma_2^n \ge \cdots \ge \gamma_l^n$, respectively. Then, for each URLLC packet in the list, the BS tests the feasibility condition between the URLLC packet and the first available eMBB users. If the feasibility conditions hold, then the BS allocates the URLLC packet over this eMBB user and update both R_e^{th} and ϕ_e, otherwise the bases station repeats the feasibility test with the next eMBB user. Noting that, more than one URLLC packet can be allocated over one eMBB user. The Boolean variable aims to reduce the time complexity of the proposed algorithm as the URLLC packets are sorted in descending order.

Algorithm 1: URLLC allocation Algorithm.

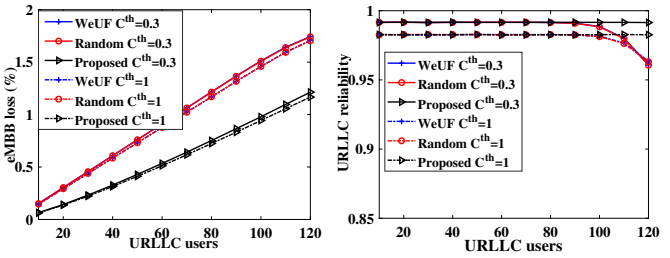
```
- Sort eMBB users in ascending order based on γ_e ;
- Sort URLLC users in descending order based on γ_l^n ;
for l = 1 → l^n do
    Boolean=0 allocating indicator variable;
    for e = 1 → |E| do
        if feasibility conditions in (14) and (15) then
            allocate URLLC packet l on the eMBB
              user e;
            Update R̂_e^th and φ_e;
            Boolean=1;
            break;
    if Boolean==0 then
        break;
```

V. Simulation Results

A. Simulation setup:

In this section, we carry out simulations to numerically evaluate the performance of the proposed superposition algorithm. We consider a wireless network which consists of one BS, 100 resource blocks, 10 eMBB users, and several URLLC users. The eMBB rate loss is assumed to be 2% of the eMBB rate, i.e., $R_e^{th} = 0.02R_e$. We consider a high URLLC load consisting of up to 120 users; accordingly, at 60 URLLC users with packet generation probability $p = 0.08$, the average arrival packets are 38400 packets per second, which is relatively high [16]. The proposed algorithm is compared with two baseline algorithms. The first algorithm works by puncturing the weakest eMBB user first (WeUF); that is, the eMBB user with the worst channel condition is punctured first until the loss threshold is reached [11]. The advantage of the WeUF algorithm is to minimize the eMBB loss by puncturing eMBB users with the lowest achievable rates. The second algorithm is to puncture at random the eMBB users as long as that satisfies the loss threshold for the punctured user. Random placement is used to allocate the URLLC load due to the optimality of random placement for the linear loss model; this follows from the fact that if eMBB resource is punctured uniformly, then the punctured resources are proportional to the bandwidth assigned to the eMBB user [6]. Simulation results are performed over 10^5 independent channel gain realizations.

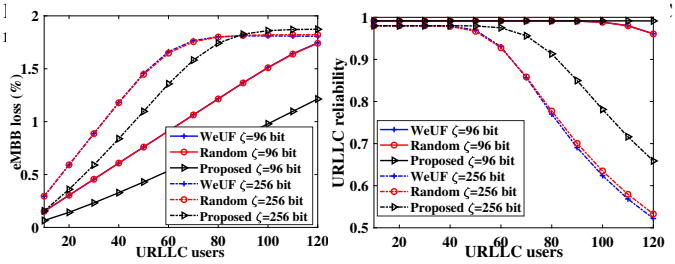
The impact of the URLLC rate threshold on both URLLC reliability and the eMBB rate loss is illustrated in Fig. 3. The figure shows that the proposed algorithm achieves better



(a) eMBB loss vs number of URLLC users.

(b) URLLC reliability vs number of URLLC users.

Fig. 3: Performance analysis of proposed algorithm with different URLLC rate threshold and $\zeta = 96$.



(a) eMBB loss vs number of URLLC users.

(b) URLLC reliability vs number of URLLC users.

Fig. 4: Performance analysis of proposed algorithm with different URLLC packet size and $C^{th} = 0.3\ bits/sec/Hz$.

eMBB rate loss and URLLC reliability compared to that of the puncturing baselines, hence, the proposed algorithm can accommodate more URLLC traffic compared to puncturing baselines. That is because the eMBB resources are entirely lost if puncturing is used, while the eMBB user adopts SIC to extract superimposed resources if superposition is used. In line with the results of [6], the figure also shows that random puncturing has slightly better performance WeUF. Moreover, Fig. (3b) shows that URLLC reliability decreases by increasing C^{th} as more URLLC packets are dropped due to URLLC rate requirement; hence, eMBB loss enhanced (see Fig. (3a)). Furthermore, increasing the URLLC load affects the URLLC reliability because there are not enough resources to allocate more URLLC packets. In other words, when the URLLC load is low the URLLC packet dropping is dominated by C^{th} while at high URLLC load the URLLC is dominated by R_e^{th}.

Similarly, Fig. (4) illustrates the effect of the URLLC packet size on both the URLLC reliability and the eMBB rate loss. As shown in Fig. (4a), when URLLC packet size ζ becomes larger the URLLC load increasing. Hence, eMBB loss keeps increasing until the eMBB rate threshold is reached. Fig. (4b) shows that URLLC reliability is inversely proportional to ζ; this due to the impact of the URLLC channel condition on the URLLC rate threshold. Furthermore, for large URLLC packet size, URLLC reliability as there significantly reduces; this due to the high URLLC traffic exceeds capacity that satisfies the eMBB QoS. In summary, the eMBB QoS may affect the URLLC reliability, so it should be jointly optimized with the eMBB rate while considering the URLLC load.

VI. Conclusions

In this paper, we proposed a low-complexity resource allocation scheme in a downlink network which consists of a single base station serving simultaneously URLLC and eMBB users. To minimize the eMBB rate loss, we formulated the allocation problem as an MINLP which is generally hard to solve. We first derived the feasibility region and the optimal solution for the case of one-to-one pairing. Then, we applied the results for the case of many-to-one pairing. Simulation results showed that the proposed algorithm achieves better URLLC reliability and eMBB rate while satisfying the QoS of the eMBB users compared to the state of the art puncturing baselines. Moreover, the proposed algorithm has low time complexity which is in order of sub-millisecond, making it an efficient tool to be used in practice. In future work, the case of many-to-many pairing will be considered while minimizing the signaling loss due to URLLC packet segmentation.

Acknowledgement

The authors acknowledge the financial support from the Natural Sciences and Engineering Research Council of Canada (NSERC), Fonds Québécois de la Recherche sur la Nature et les Technologies (FQRNT) and from Concordia University.

References

[1] M. Bennis, M. Debbah, and H. V. Poor, "Ultrareliable and Low-Latency Wireless Communication: Tail, Risk, and Scale," *Proceedings of the IEEE*, vol. 106, no. 10, pp. 1834–1853, Oct 2018.

[2] W. Saad, M. Bennis, and M. Chen, "A Vision of 6G Wireless Systems: Applications, Trends, Technologies, and Open Research Problems," *IEEE Network*, vol. 34, no. 3, pp. 134–142, 2020.

[3] K. B. L. *et al.*, "The Roadmap to 6G: AI Empowered Wireless Networks," *IEEE Commun. Mag.*, vol. 57, no. 8, pp. 84–90, 2019.

[4] P. *et al.*, "Extreme URLLC: Vision, challenges, and key enablers," *arXiv preprint arXiv:2001.09683*, 2020.

[5] H. Ji *et al.*, "Ultra-Reliable and Low-Latency Communications in 5G Downlink: Physical Layer Aspects," *IEEE Wireless Commun.*, vol. 25, no. 3, pp. 124–130, JUNE 2018.

[6] A. Anand, G. De Veciana, and S. Shakkottai, "Joint Scheduling of URLLC and eMBB Traffic in 5G Wireless Networks," in *IEEE Conf. INFOCOM 2018*, Apr 2018, pp. 1970–1978.

[7] A. Karimi *et al.*, "Efficient Low Complexity Packet Scheduling Algorithm for Mixed URLLC and eMBB Traffic in 5G," in *2019 IEEE 89th Veh. Technol. Conf. (VTC2019-Spring)*, Apr 2019, pp. 1–6.

[8] M. Alsenwi *et al.*, "eMBB-URLLC Resource Slicing: A Risk-Sensitive Approach," *IEEE Commun. Lett.*, vol. 23, no. 4, pp. 740–743, Apr 2019.

[9] M. Alsenwi *et al.*, "Intelligent Resource Slicing for eMBB and URLLC Coexistence in 5G and Beyond: A Deep Reinforcement Learning Based Approach," *arXiv preprint arXiv:2003.07651*, 2020.

[10] A. Manzoor *et al.*, "Contract-based Scheduling of URLLC Packets in Incumbent EMBB Traffic," *arXiv preprint arXiv:2003.11176*, 2020.

[11] K. I. Pedersen *et al.*, "Punctured Scheduling for Critical Low Latency Data on a Shared Channel with Mobile Broadband," in *2017 IEEE 86th Veh. Technol. Conf. (VTC-Fall)*. IEEE, 2017, pp. 1–6.

[12] Y. Saito *et al.*, "Non-Orthogonal Multiple Access (NOMA) for Cellular Future Radio Access," in *2013 IEEE VTC Conf.*, 2013, pp. 1–5.

[13] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel Coding Rate in the Finite Blocklength Regime," *IEEE Trans. on Information Theory*, vol. 56, no. 5, pp. 2307–2359, 2010.

[14] J. Cao *et al.*, "Joint Block Length and Pilot Length Optimization for URLLC in the Finite Block Length Regime," in *IEEE GLOBECOM*, 2019, pp. 1–6.

[15] S. Ebbesen, P. Kiwitz, and L. Guzzella, "A Generic Particle Swarm Optimization Matlab Function," in *2012 American Control Conf. (ACC)*, 2012, pp. 1519–1524.

[16] D. Maaz, A. Galindo-Serrano, and S. E. Elayoubi, "URLLC User Plane Latency Performance in New Radio," in *2018 25th Int. Conf. on Telecommun. (ICT)*. IEEE, 2018, pp. 225–229.