eMBB-URLLC Resource Slicing: A Risk-Sensitive Approach

Madyan Alsenwi, Nguyen H. Tran, Senior Member, IEEE, Mehdi Bennis, Senior Member, IEEE, Anupam Kumar Bairagi, Member, IEEE, and Choong Seon Hong, Senior Member, IEEE

Abstract—Ultra Reliable Low Latency Communication (URLLC) is a 5G New Radio (NR) application that requires strict reliability and latency. URLLC traffic is usually scheduled on top of the ongoing enhanced Mobile Broadband (eMBB) transmissions (i.e., puncturing the current eMBB transmission) and cannot be queued due to its hard latency requirements. In this letter, we propose a risk-sensitive based formulation to allocate resources to the incoming URLLC traffic while minimizing the risk of the eMBB transmission (i.e., protecting the eMBB users with low data rate) and ensuring URLLC reliability. Specifically, the Conditional Value at Risk (CVaR) is introduced as a risk measure for eMBB transmission. Moreover, the reliability constraint of URLLC is formulated as a chance constraint and relaxed based on Markov’s inequality. We decompose the formulated problem into two subproblems in order to transform it into a convex form and then alternatively solve them until convergence. Simulation results show that the proposed approach allocates resources to the incoming URLLC traffic efficiently while satisfying the reliability of both eMBB and URLLC.

Index Terms—5G NR, URLLC, eMBB, latency, reliability, resource slicing, risk-sensitive, CVaR.

I. INTRODUCTION

THE upcoming Fifth Generation (5G) New Radio (NR) is designed to support three major types of traffic: enhanced Mobile Broad Band (eMBB), Massive Machine Type Communications (mMTC), and Ultra Reliable Low Latency Communications (URLLC). While eMBB is an extension of the LTE-Advanced service whose objective is to maximize the peak data rate, mMTC is designed to support a large number of Internet of Things (IoT) devices sending small data sporadically during the active phase only. On the other hand, URLLC is designed to support services that require high level of reliability and low latency. According to the Third Generation Partnership Project (3GPP), the main objective of URLLC is to minimize the latency down to 1ms while ensuring packet error rates of less than $10^{-5}$ [1]. These requirements are critical for applications such as industrial IoT, autonomous vehicles, and virtual reality [2, 3].

The arrival URLLC traffic is immediately transmitted due to its hard latency requirements which may overlap onto previously allocated eMBB transmissions. Allocating resources to the critical URLLC traffic based on a formulation that aims to maximize the total average data rate of eMBB users hinders eMBB users with low data rate and protects higher data rate users. Therefore, we depart from the classical average-based formulation and instead capture the tail of the rate distribution. Considering this risk for eMBB transmission protects low data rate eMBB users and ensures reliable eMBB transmission.

Recently, studies focusing on URLLC have gained attention in both academia and industry. The authors in [3] discuss the principles for achieving URLLC and describe several building blocks of framing, use of diversity, and access topology. Authors in [5] consider a linear model, convex model, and threshold model for eMBB data rate loss associated with URLLC traffic. Authors in [4] propose a punctured scheduling approach for transmission of low latency communication (LLC) traffic multiplexed on a shared channel with eMBB. In this work, downlink resource slicing for URLLC and eMBB traffics based on puncturing is considered. In summary, our contributions are as follows:

- We first formulate a risk-sensitive optimization problem to find the probability of each eMBB user being punctured by URLLC downlink traffic. In contrast to the classical average-based formulation, the risk-sensitive formulation captures the tail of eMBB rate distribution ensuring reliability by protecting users with low data rate. We use the Conditional Value at Risk (CVaR) as a risk measure. Furthermore, we formulate the reliability constraint of URLLC as a chance constraint. In addition, RBs are allocated to eMBB users based on formulation that guarantees proportional fairness.

- Due to the non-convexity of the formulated problem, we decompose it into two subproblems: 1) eMBB users scheduling and 2) URLLC placement problem. The eMBB users scheduling problem is an integer programming problem. Therefore, we relax it to a convex optimization problem whose solution is within a constant approximation from the optimal. Furthermore, the Markov’s inequality is leveraged to relax the URLLC chance constraint of the URLLC placement problem into a linear form. Therefore, the resulting URLLC placement problem is a transformed convex optimization problem for a given eMBB users scheduling which can be solved efficiently by the Base Station (BS). The two problems are then alternatively maximized until convergence.

- Simulation results show that the proposed approach allocates resources to the incoming URLLC traffic while ensuring the reliability of both eMBB and URLLC. The results show that our proposed approach keeps the eMBB reliability higher than 90% for different URLLC traffics. Moreover, the results show the tradeoff between eMBB data rate and URLLC reliability.
be approximated based on the Shannon capacity model as follows:
\[ R_u = \varphi_u \left( 1 - \omega_u \times \frac{L}{L_{\text{max}}} \right) \log_2 \left( 1 + \frac{P_u g_u}{N_0} \right), \]
where \( L = \sum_{m=1}^{M} L_m \) follows the binomial distribution with parameters \( M \) and \( p \), \( L_{\text{max}} \) is the maximum URLLC traffic that can be served at a time slot (i.e., \( L_{\text{max}} \) equals to the system capacity), \( P_u \) and \( g_u \) are the transmission power and channel gain of the user \( u \) respectively, and \( N_0 \) is the noise power. The term \( (\omega_u \times L/L_{\text{max}}) \) is an approximation of the punctured resources of eMBB user \( u \) by URLLC traffic, where \( L/L_{\text{max}} \) is the normalized URLLC traffic and \( \varphi_u \) represents the total resources of eMBB user \( u \). URLLC traffic exceeding system capacity are blocked thus \( L \leq L_{\text{max}} \) holds almost surely.

\[ \text{VaR}_\alpha(R) = \arg\inf \{ \gamma : P(R > \gamma) \leq \alpha \}, \]
where \( R = \sum_{u \in \mathcal{U}} R_u \), and \( \alpha \in (0,1) \). The CVaR function is defined as the expectation of the \( \alpha \) fraction of the worst outcomes of \( R \):
\[ \text{CVaR}_\alpha(R) = \mathbb{E}[R | R > \text{VaR}_\alpha(R)]. \]
Moreover, we have that \[ \phi_\alpha(R, \gamma) := \gamma + \frac{1}{1 - \alpha} \mathbb{E}[ (R - \gamma)^+] , \]
where \((x)^+ = \max(0, x)\), and the CVaR of the random variable \( R \) can be determined as
\[ \text{CVaR}_\alpha(R) = \min_{\gamma \geq 0} \phi_\alpha(R, \gamma). \]
As \( \mathbb{E}[L] = Mp \), we can rewrite the \( \mathbb{E}[R] \) in (4) as follows:
\[ \mathbb{E}[R] = \sum_{u \in \mathcal{U}} \varphi_u \left( 1 - \omega_u \times \frac{Mp}{L_{\text{max}}} \right) \log_2 \left( 1 + \frac{P_u g_u}{N_0} \right). \]

Let \( C = \{1, 2, ..., C\} \) be the set of all URLLC users, \( R_c \) is the data rate of a URLLC user \( c \), and \( R_{urlc} = \sum_{c \in C} R_c \).
represents the data rate of all URLLC users. Therefore, the outage probability of URLLC is given as:

\[ P(E) = Pr[R_{urllc} \leq L], \tag{7} \]

where \( R_{urllc} \) is given by:

\[ R_{urllc}(\theta) = \sum_{c \in C} \sum_{u \in U} \frac{\theta_u}{C} \log_2 \left( 1 + \frac{P_c R_{urllc}}{N_0} \right). \tag{8} \]

In (8), \( \theta_u = (\phi_u / \omega_u \max_{\omega_u \in \Omega} \) represents the punctured resources from the eMBB user \( u \) to URLLC traffic, \( P_c \) and \( g_c \) are the transmission power and channel gain of URLLC user \( c \), respectively. Here, we consider that the total punctured resources to URLLC traffic is divided equally between the URLLC users.

RBs are allocated to eMBB users at the beginning of each time slot by solving an optimization problem that guarantees proportional fairness. The proportional fair resource allocation is modeled as the sum-log formulation to ensure proportional fairness. The proportional fair resource allocation problem seeks both the optimum resource allocation matrix \( I \) and the other parameters remain constant for sufficiently long RBs are allocated to eMBB users at the beginning of each time slot by solving an optimization problem that guarantees proportional fairness. The proportional fair resource allocation is modeled as the sum-log formulation to ensure proportional fairness. The proportional fair resource allocation problem seeks both the optimum resource allocation matrix \( I \) and the other parameters remain constant for sufficiently long

\[ \text{maximize} \quad \sum_{u \in U} \log \left( E[R_u] \right) - \beta \left( \text{CVaR}_u (\gamma) \right) \tag{9a} \]

subject to

\[ \text{Pr}[R_{urllc} \leq L] \leq \epsilon, \tag{9b} \]

\[ \sum_{u \in U} I_{b,u} \leq 1, \quad \forall b \in \mathcal{B}, \tag{9c} \]

\[ I_{b,u} \in \{0, 1\}, \quad \forall u \in \mathcal{U} \text{ and } b \in \mathcal{B}, \tag{9d} \]

\[ 0 \leq \omega_u \leq 1, \quad \forall u \in \mathcal{U}, \tag{9e} \]

where \( \beta \in [0, 1] \) is the weight of the CVaR function, and \( \epsilon \) denotes the reliability of URLLC and takes a small value. The optimization problem seeks both the optimum resource allocation matrix \( I^* \) for eMBB users and the optimum URLLC placement weight vector \( \omega^* \). The objective function (9a) is formulated based on a sum-log formulation to ensure proportional fairness when allocating the resources to eMBB users. Furthermore, the CVaR function minimizes the risk of eMBB users when allocating resources to URLLC traffic.

III. PROPOSED SOLUTION

The original problem (9) is a mixed-integer nonlinear programming (MINLP). To find a global optimum solution, we need to search the space of feasible URLLC placement weights with all possible combinations of eMBB user scheduling. This may require exponential-complexity to solve. To solve this problem efficiently, we decompose the original problem into two subproblems: 1) eMBB users scheduling, and 2) URLLC placement problem.

A. Resource Allocation for eMBB users

For any fixed feasible URLLC weight vector \( \omega \), the original problem (9) can be presented as follows:

\[ \text{maximize} \quad I \sum_{u \in U} \log \left( E[R_u] \right) - \beta \left( \text{CVaR}_u (\gamma) \right) \tag{10a} \]

subject to

\[ \sum_{u \in U} I_{b,u} \leq 1, \quad \forall b \in \mathcal{B}, \tag{10b} \]

\[ I_{b,u} \in \{0, 1\}, \quad \forall u \in \mathcal{U} \text{ and } b \in \mathcal{B}. \tag{10c} \]

B. URLLC Scheduler for a Given eMBB User Scheduling Matrix

For any given eMBB user scheduling \( I \), the original problem (9) can be reduced to the following URLLC placement strategy problem:

\[ \text{maximize} \quad \sum_{u \in U} \log \left( E[R_u(\omega)] \right) - \beta \left( \gamma + \frac{1}{1 - \alpha} \left( \text{Pr}[R_{urllc}(\omega)] - \gamma \right) \right) \tag{12a} \]

subject to

\[ \text{Pr}[R_{urllc}(\omega) \leq L] \leq \epsilon, \quad 0 \leq \omega_u \leq 1, \quad \forall u \in \mathcal{U}. \tag{12b} \]

We use the Markov’s Inequality to represent the chance constraint (12b) as a linear constraint:

\[ \text{Pr}[R_{urllc} \leq L] \leq \frac{\text{E}[L]}{R_{urllc}}. \tag{13} \]

Furthermore, we introduce variable \( z \) to replace \( (\text{Pr}[R_{urllc}(\omega)] - \gamma)^+ \). This is achieved by imposing the constraints \( z \geq \text{Pr}[R_{urllc}(\omega)] - \gamma \) and \( z \geq 0 \):

\[ \text{maximize} \quad \omega, z, \gamma \sum_{u \in U} \log \left( E[R_u(\omega)] \right) - \beta \left( \gamma + z \right) \frac{1}{1 - \alpha} \tag{14a} \]

subject to

\[ z \geq \text{Pr}[R_{urllc}(\omega)] - \gamma, \quad 0 \leq \omega_u \leq 1, \quad \forall u \in \mathcal{U}. \tag{14b} \]

For any given eMBB user scheduling, the above problem is a convex optimization problem which can be solved efficiently by the BS. The BS only needs to estimate the wireless channel gains of eMBB users that are time varying in each time slot and the other parameters remain constant for sufficiently long

\[ \text{Algorithm 1 Resource Allocation Strategy for eMBB and URLLC Traffic} \]

1: URLLC weight vector initialization
2: repeat:
3: eMBB users scheduling (11)
4: URLLC placement problem (14)
5: until \( I \) and \( \omega \) converge or max # of iterations is reached

The integer programming problem (10) can be relaxed to a convex optimization problem whose solution is within a constant approximation from the optimal. Then, the fractional rounding is used to get a solution to the original integer problem. The randomization of \( R_u \) comes from the URLLC part only (i.e., we consider that the BS can estimate the channel gain of all users). Therefore, considering only the RBs allocation to eMBB users problem leads to a deterministic variable of \( R_u \). In this case, the CVaR can be removed from the objective function. Accordingly, the optimization problem (10) can be approximated as follows:

\[ \text{maximize} \quad I \sum_{u \in U} \log \left( E[R_u] \right) - \beta \left( \text{CVaR}_u (\gamma) \right) \tag{11a} \]

subject to

\[ \sum_{u \in U} I_{b,u} \leq 1, \quad \forall b \in \mathcal{B}, \tag{11b} \]

\[ 0 \leq I_{b,u} \leq 1, \quad \forall u \in \mathcal{U} \text{ and } b \in \mathcal{B}. \tag{11c} \]

The convex optimization problem (11) can be solved by applying the Karush-Kuhn-Tucker (KKT) conditions.
increasing $p$ means having more URLLC traffic and this leads to more impacting on the eMBB transmission. We also note that Baseline 2 gives higher eMBB data rate since its objective is to maximize the sum-rate of eMBB transmissions without considering its reliability.

Finally, Fig. 3 shows the data rate of eMBB users and the distributed URLLC traffic over eMBB users. As shown in Fig. 3 (a), the proposed risk-sensitive based approach gives the eMBB users with high data rates higher probability to be punctured by URLLC traffic while protecting the eMBB users with bad channel conditions. However, setting $\beta = 0$ (Baseline 1) distributes the incoming URLLC traffic equally among all eMBB users since it aims at a proportional-fair as shown in Fig. 3(b). Furthermore, Baseline 2 gives the eMBB users with low data rate higher probability to be punctured by URLLC traffic since its objective is to maximize the sum eMBB data rate as shown in Fig. 3(c).

V. CONCLUSIONS

In this work, we studied the problem of dynamic multiplexing of URLLC traffic by puncturing eMBB resources. The resource scheduling problem is formulated as an optimization problem that aims to maximize the total eMBB data rate while considering the risk of eMBB using the CVaR function as a risk measure. The results have shown that the proposed algorithm protects eMBB users with low data rate by allocating more URLLC traffic to eMBB users with higher data rate.

REFERENCES

[1] 3GPP TSG RAN WG1 95, Tech. Rep., November 2018.
[2] M. Bennis, M. Debbah, and H. V. Poor, “Ultra-reliable and low-latency wireless communication: Tail, risk and scale,” arXiv preprint arXiv:1801.01270, 2018.
[3] P. Popovski, I. I. Nielsen, C. Stefanovic, E. de Carvalho, E. Strom, K. F. Trillingsgaard, A.-S. Bana, D. M. Kim, R. Kotaba, J. Park, and R. B. Sorensen, “Wireless access for ultra-reliable low-latency communication: Principles and building blocks,” IEEE Network, vol. 32, no. 2, pp. 16–23, mar. 2018.
[4] K. I. Pedersen, G. Pocovi, J. Steiner, and S. R. Khoorsarmid, “Punctured scheduling for critical low latency data on a shared channel with mobile broadband,” IEEE Vehicular Technology Conference, vol. 2017-Septe, no. 2, pp. 1–6, 2018.
[5] A. Anand, G. de Veciana, and S. Shakkottai, “Joint Scheduling of URLLC and eMBB Traffic in 5G Wireless Networks,” IEEE INFOCOM 2018-IEEE Conference on Computer Communications, IEEE, pp. 1–12, 2018. [Online]. Available: http://arxiv.org/abs/1712.05344
[6] R. T. Rockafellar and S. Uryasev, “Optimization of conditional value-at-risk,” The Journal of Risk, vol. 2, no. 3, pp. 21–41, 2000.
[7] T. Girici, C. Zhu, J. R. Agre, and A. Ephremides, “Proportional fair scheduling algorithm in OFDMA-based wireless systems with QoS constraints,” Journal of Communications and Networks, vol. 12, no. 1, pp. 30–42, feb 2010.