Ensuring Quality-of-Service (QoS) in eMBB-URLLC Co-existence Scenario

Yan Kyaw Tun and Choong Seon Hong
Department of Computer Science and Engineering, Kyung Hee University, Yongin-si, Gyeonggi-do 17104, Korea
email:{ykyawtun7, cshong}@khu.ac.kr

Abstract

In this work, we propose a resource scheduling optimization problem to ensure quality-of-service (QoS) in a dynamic multiplexing scenario of URLLC (Ultra-reliable low latency communication) and eMBB (Enhanced mobile broadband) services. The services requiring high bandwidth such as augmented reality and video streaming are classified as eMBB traffic, whereas services like remote surgery and autonomous driving that demand sub-millisecond latency with minimum error rates are classified as URLLC traffic in 5G New Radio (NR). We characterized these services for a dynamic multiplexing scenario, where we target to ensure the stringent requirements of URLLC traffics while guaranteeing the data rate of eMBB users. To that end, we formulate a resource scheduling optimization problem with multiplexed eMBB-URLLC services. Here, we consider a puncturing technique that allows URLLC traffic to schedule over the ongoing eMBB transmissions. Next, we investigate the problem of Resource Blocks (RBs) allocation to eMBB users with a 2-Dimensions Hopfield Neural Networks (2D-HNN) formulation. Simulation results show the efficacy of our proposed method, in terms of fairness and achievable data rate.

Keywords – *5G New Radio (NR), eMBB, URLLC, resource scheduling, Neural Networks.*

I. Introduction

The upcoming 5G networks expects unprecedented surge of heterogeneous cellular services. Specifically, the services requiring stringent latency requirements, and those demanding high data rate will dominate the upcoming mobile networks. However, in the existing network, it is almost impossible to ensure both of these requirements at the same time [1]. This is due to the fact that the current network architecture primarily focuses on maximizing the overall network throughput, using long packets. In doing so, to ensure ultra-reliability, a way would be to use short packets [2]; however, it dramatically reduces the data rate. Thus, the challenge to efficiently manage radio resources for fulfilling QoS requirements of the these diverse upcoming services remains.

In 5G New Radio (NR), new features classifying these peculiar services are defined based on their requirements [3]. The services are characterized as: (i) enhanced Mobile Broad Band (eMBB), (ii) massive Machine Type Communications (mMTC), and (iii) Ultra Reliable Low Latency Communications (URLLC). The services requiring high bandwidth such as augmented reality and video streaming are classified as eMBB traffic. Basically, eMBB is characterize like internet access service, similar to an extension to Long Term Evolution-Advanced (LTE-A) [4]. The services characterizing sporadic nature of traffic, such as with Internet-of-Things (IoT), in particular, like sensing and monitoring services, is defined as mMTC. It can be considered as narrow-band internet access where the nodes are active for a short time interval. Similarly, services like remote surgery and autonomous driving that demand sub-millisecond latency with minimum error rates are classified as URLLC traffic. For an example, reliability is defined as $(1-10^{-5})$ success probability while a user transmits a Protocol Data Unit (PDU) of 32 bytes within 1ms [5]. This is defined as QoS requirements of URLLC by the current 3GPP standards.

In this work, we employ puncturing techniques, which is one of the approach to deal with the stringent requirements (i.e., in terms of latency and reliability) for spectrum management in 5G NR to satisfy URLLC [6]. This method is concurrent to the standards set by 3GPP report, where URLLC traffic needs to be immediately transmitted for meeting their QoS requirements [7]. We consider the coexistence of eMBB-URLLC traffic, where under the puncturing mechanism, the 5G NodeB (gNB) will drop ongoing eMBB transmissions to satisfy QoS requirements of URLLC traffic in the upcoming time slot. The dropped eMBB users because of puncturing will be rescheduled [8], [9], [10]. To that end, we formulated a resource scheduling optimization problem with multiplexed eMBB-URLLC services. We adopted our proposed 2-Dimensional Hopfield Neural Network (2D-HNN) solution [11] to solve the formulated problem. In doing so, we use Cumulative Distribution Function (CDF) of the random URLLC traffic to relax the chance constraint to a deterministic linear constraint in the optimization problem.

II. Problem Formulation

We sequentially tackle the problem of resource scheduling and allocation in this section. In doing so, we first define the RB allocation formulation to derive an eMBB scheduler. Next, we consider the multiplexing scenario, where we consider the QoS constraint and latency requirements of both eMBB-URLLC services. Here, a neural network based eMBB users scheduling strategy is adopted, following our earlier work [11]. Furthermore, also relax the chance constraint to derive solutions for the scheduling problem.

We consider $K \in \mathcal{K}$ eMBB users available within a time-slot T and a set of URLLC nodes $\mathcal{U} \subseteq \mathcal{K}$ requesting services in frequency-time slot. B is the effective bandwidth in each slot, which is further divided into mini slots (t_m) such that resource $f_\eta(t_m) = B/\eta$, for $\eta > 0$. Furthermore, $f_\eta(t_m)$ is quantize into N levels. Then, $x_{k(t_m),N} \in \{0,1\}$ denotes the RB association variable such that $f_\eta(t_m) = x_k^N(t_m) \cdot N$, for each associated user k.

A. Resource Blocks Allocation to eMBB Traffics

The instantaneous rate for an eMBB user k at time slot T is given as

$$R_k^e = \sum\nolimits_{N\in\mathcal{N}} f_\eta(t_m) x_k^N(t_m) x_{k(t_m),N} \log_2\left(1 + \frac{\rho_k |G_k|^2}{N_o}\right), \quad \forall N \in \mathcal{N}, \tag{1}$$

where $f_\eta(t_m)x_k^N(t_m)x_{k(t_m),N}$ is the total spectrum allocated, ρ_k is the transmission power, $|G_k|^2$ is the channel gain between user k and the base station, and N_o is the noise power. Then, following (1), the rate of all eMBB users in time slot T is

$$R_k^e(T) = \sum\nolimits_{k\in\mathcal{K}} \sum\nolimits_{N\in\mathcal{N}} f_\eta(t_m) x_k^N(t_m) x_{k(t_m),N} \cdot \log_2\left(1 + \frac{\rho_k |G_k|^2}{N_o}\right), \ \forall k \in \mathcal{K}, N \in \mathcal{N} \tag{2}$$

such that $\sum_{k\in\mathcal{K}} f_\eta(t_m) x_k^N(t_m) x_{k(t_m),N} \leq B$.

In this regards, we can define the average data rate for user k up to time t as

$$\tilde{R}_k^e(t) = \zeta \tilde{R}_k^e(t-1) + (1-\zeta) R_k^e(t) \tag{3}$$

where $\zeta \in [0,1]$. Therefore, the optimization problem for eMBB scheduler can be formulated as

$$\begin{aligned} \underset{\mathrm{x}}{\text{Max}} \quad & \sum\nolimits_{k\in\mathcal{K}} \frac{R_k^e(T)}{[\tilde{R}_k^e(T)]^\alpha} \\ \text{s.t} \quad & C_1 : \sum\nolimits_{N\in\mathcal{N}} x_{k(t_m),N} \leq 1, \ \forall k \in \mathcal{K} \\ & C_2 : x_{k(t_m),N} \in \{0,1\}, \ \forall k \in \mathcal{K}, N \in \mathcal{N}. \end{aligned} \tag{4}$$

where the constraint (4)C_1 ensures that each RB is allocated to only one user at a time. The solution of (4) is the allocation matrix x with each element defined as

$$x_{k(t_m),N} = \begin{cases} 1, & \text{if } N \in \mathcal{N}_k; \\ 0, & \text{Otherwise}, \end{cases} \tag{5}$$

where $\mathcal{N}_k$ is the set of all RBs allocated to the eMBB user k. For shorthand representation, we will use $x_{k,N}$ for $x_{k(t_m),N}$ hereafter.

B. Resource Scheduling Mechanism for eMBB-URLLC Traffics

In a scenario of dynamic multiplexing between eMBB-URLLC traffics, we assume that the impact of eMBB traffic to punctured resources by URLLC traffic is proportional [12]. Consider η_u^k is the level of punctured RBs within a time slot T of eMBB user k i.e., the impact on the data rate of eMBB users.

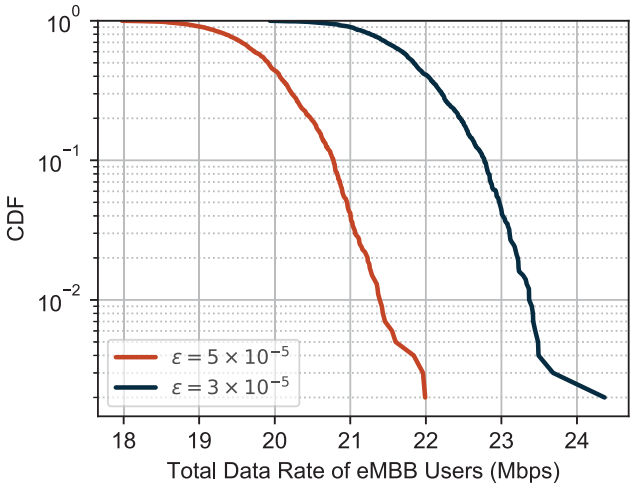

Fig. 1. CDF of the total data rate at $\alpha = 1$.

Let the arriving URLLC traffic load during time t is characterized by the random variable $X(t)$. Then, the corresponding outage probability of URLLC traffic is defined as

$$P(O) = P(R_T^u < X(t)) \tag{6}$$

where R_T^u is the instantaneous rate of URLLC traffic i.e.,

$$P(O) = P\left[\sum\nolimits_{k\in\mathcal{K}} \eta_u^k \log_2\left(1 + \frac{\rho_u |G_u|^2}{N_o}\right) < X(t)\right]. \tag{7}$$

Therefore, the proposed scheduler aims at maximizing the total data rate of eMBB users while satisfying the latency stringent constraint of URLLC traffic as follows:

$$\begin{aligned} \underset{\eta_u}{\text{Max}} \quad & \sum\nolimits_{k\in\mathcal{K}} \sum\nolimits_{N\in\mathcal{N}} \left(f_\eta(t_m) x_k^N(t_m) x_{k,N} - \eta_u^k\right) \\ & \cdot \log_2\left(1 + \frac{\rho_k |G_k|^2}{N_o}\right) \\ \text{s.t} \quad & C_1 : P(R_T^u < X) \leq \epsilon \\ & C_2 : \sum\nolimits_{k\in\mathcal{K}} \eta_u^k \leq f_\eta(t_m) x_k^N(t_m) x_{k,N}, \ \forall k \in \mathcal{K} \end{aligned} \tag{8}$$

where (8)C_1 characterizes the maximum outage probability of the URLLC traffic, namely the *reliability level* with the probability value ϵ. Constraint (8)C_2 ensures that the proportion of resources to the URLLC load is no more than the allocated resources for eMBB users. In order to obtain a close form solution for the optimization problem, we first need to relax the chance constraint (8)C_1. Then, we solve the problem by using 2-Dimensions Hopefield Neural Networks (2D-HNN).

III. Simulation Results

We evaluate the performance of our proposed solution approach in terms of achieved data rate and fairness. For this, we consider 10 eMBB users with different channel states. In each time slot, we consider that 100 RBs are available, and each RB is 180 kHz. Note that 5G NR permits a large number configuration varying from 15kHz to 480kHz. We will present them accordingly in this section.

In Fig. 1, for the different values of reliability metrics ϵ at $\alpha = 1$, we evaluate the CDF of the total data rate of all eMBB

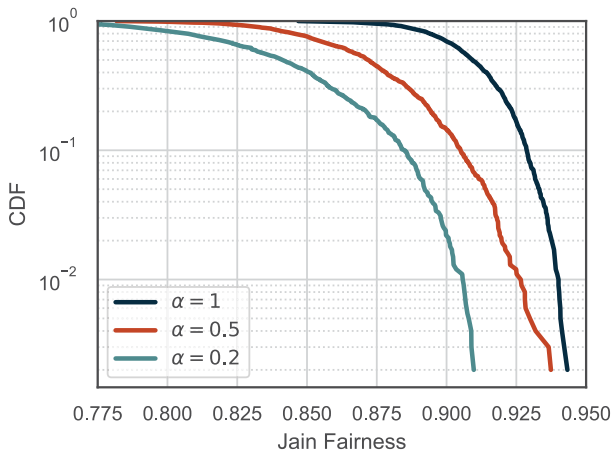

Fig. 2. Comparison of Fairness.

users. It is observed that a higher value of ϵ will limit the puncturing to satisfy URLLC reliability requirements. Thus, we see the increase in the overall sum rate for eMBB users. On the contrary, smaller values of ϵ means that we provide higher priority to the URLLC reliability constraint. Consequently, we observe the drop in the data rate of eMBB users as more resources are punctured.

Next, we use different values α and the reliability level of URLLC load ϵ, and evaluate the performance of the proposed mechanism following hopfield neural network solutions, as in [11]. We calculate the long-term data rate of all eMBB users and fairness among them for different parametric values α and ϵ. We use Jain's fairness index [13] to evaluate the fairness amongst the users which is given as

$$f(T_1, ..., T_z) = \frac{(\sum_{i=1}^{z} T_i)^2}{z \sum_{i=1}^{z} T_i^2}. \quad (9)$$

The value of Jain's fairness index lies within the interval [0, 1], i.e., around 0.95 for $\alpha = 1$, i.e, the system's fairness index is 0.95 which means that it is 95% fair).

IV. Conclusion

In this paper, we have studied a resource scheduling mechanism to ensure quality-of-service (QoS) in a dynamic multiplexing scenario of URLLC (Ultra-reliable low latency communication) and eMBB (Enhanced mobile broadband) services. Here, the scheduled eMBB traffic is punctured by URLLC traffic to fulfill its stringent latency requirements. In doing so, we have defined the RB allocation formulation to derive an eMBB scheduler that maximizes the overall sum-rate of eMBB users. Next, we have considered the multiplexing scenario, where we have the QoS constraint and latency requirements of both eMBB-URLLC services. We have solved the resource allocation problem with a modified 2D-HNN. In doing so, we have relaxed the chance constraint problem to a deterministic constraint using the CDF of the arrival URLLC traffic. Using the comparison results, we have demonstrated the efficacy of our proposed mechanism, where we have highlighted the impact of QoS constraints upon overall sum-rate of the network.

ACKNOWLEDGEMENT

This work was supported by Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2019-0-01287, Evolvable Deep Learning Model Generation Platform for Edge Computing) *Dr. CS Hong is the corresponding author.

References

[1] B. Soret, P. Mogensen, K. I. Pedersen, and M. C. Aguayo-Torres, "Fundamental tradeoffs among reliability, latency and throughput in cellular networks," in *Globecom Workshops (GC Wkshps), 2014*. IEEE, 2014, pp. 1391–1396.

[2] 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, 2010.

[3] "3gpp tsg ran wg1 88, tech. rep., february 2017," Tech. Rep.

[4] C. Hoymann, D. Astely, M. Stattin, G. Wikstrom, J.-F. Cheng, A. Hoglund, M. Frenne, R. Blasco, J. Huschke, and F. Gunnarsson, "Lte release 14 outlook," *IEEE Communications Magazine*, vol. 54, no. 6, pp. 44–49, 2016.

[5] P. Popovski, J. J. Nielsen, C. Stefanovic, E. De Carvalho, E. Strom, K. F. Trillingsgaard, A.-S. Bana, D. M. Kim, R. Kotaba, J. Park *et al.*, "Wireless access for ultra-reliable low-latency communication: Principles and building blocks," *Ieee Network*, vol. 32, no. 2, pp. 16–23, 2018.

[6] K. I. Pedersen, G. Pocovi, J. Steiner, and S. R. Khosravirad, "Punctured scheduling for critical low latency data on a shared channel with mobile broadband," in *Vehicular Technology Conference (VTC-Fall), 2017 IEEE 86th*. IEEE, 2017, pp. 1–6.

[7] Tech. Rep., study on New Radio Access Technology Physical Layer Aspects (Release 14).

[8] M. Alsenwi, N. H. Tran, M. Bennis, A. K. Bairagi, and C. S. Hong, "embb-urllc resource slicing: A risk-sensitive approach," *IEEE Communications Letters*, vol. 23, no. 4, pp. 740–743, 2019.

[9] S. R. Pandey, M. Alsenwi, Y. K. Tun, and C. S. Hong, "A downlink resource scheduling strategy for urllc traffic," in *2019 IEEE International Conference on Big Data and Smart Computing (BigComp)*. IEEE, 2019, pp. 1–6.

[10] M. Alsenwi, S. R. Pandey, Y. K. Tun, K. T. Kim, and C. S. Hong, "A chance constrained based formulation for dynamic multiplexing of embb-urllc traffics in 5g new radio," in *The International Conference on Information Networking (ICOIN 2019)*. IEEE, 2019, Kuala Lumpur, Malaysia.

[11] M. Alsenwi, I. Yaqoob, S. R. Pandey, Y. K. Tun, A. K. Bairagi, L.-w. Kim, and C. S. Hong, "Towards coexistence of cellular and wifi networks in unlicensed spectrum: A neural networks based approach," *IEEE Access*, vol. 7, pp. 110 023–110 034, 2019.

[12] A. Anand, G. de Veciana, and S. Shakkottai, "Joint scheduling of urllc and embb traffic in 5g wireless networks," *arXiv preprint arXiv:1712.05344*, 2017.

[13] R. K. Jain, D.-M. W. Chiu, and W. R. Hawe, "A quantitative measure of fairness and discrimination," *Eastern Research Laboratory, Digital Equipment Corporation, Hudson, MA*, 1984.