Uplink Scheduling for Mixed Grant-Based eMBB and Grant-Free URLLC Traffic in 5G Networks

Mohamed W. Nomeir, Yasser Gadallah, and Karim G. Seddik

Department of Electronics and Communications Engineering, American University in Cairo, Cairo, Egypt 11835
mohamednomeir@aucegypt.edu, ygadallah@ieee.org, kseddik@aucegypt.edu

***Abstract*—Scheduling in 5G networks is a challenging task due to the heterogeneous Quality of Service (QoS) requirements of traffic sources. In this paper, we consider the problem of uplink scheduling in 5G networks for mixed traffic that includes Ultra-Reliable Low Latency Communications (URLLC) devices and enhanced Mobile Broad-Band (eMBB) users. For this purpose, a mathematical model for Grant Free (GF) services is derived for the k-repetitions Hybrid Automatic Repeat reQuest (HARQ). We formulate the scheduling problem as a mixed-integer non-linear programming optimization problem. We introduce a complete system model that includes grant-free and grant-based subsystems. We then introduce our proposed solution to the scheduling problem that addresses the two traffic types. Different scheduling techniques are then compared and a performance upper bound is added as a reference. The results show that the proposed technique provides near-optimal results and outperforms other scheduling techniques with a significant complexity reduction.**

***Index Terms*—eMBB Grant Based scheduling; Genetic Algorithm; k-repetitions HARQ; Uplink scheduling; URLLC Grant Free scheduling.**

I. Introduction

The fifth-generation (5G) new radio (NR) supports heterogeneous services with different key requirements [1], [2]. Two of the main services are the enhanced Mobile Broad-Band (eMBB) communications and Ultra-Reliable Low Latency Communications (URLLC), each has different requirements [3]. The eMBB users require high data rates while moderate packet latency is acceptable. On the other hand, URLLC nodes, also known as Critical Machine Type Communications Devices (C-MTCDs), have strict low latency requirements and their payload is generally small in size.

In the Uplink direction, there are three main types of scheduling in 5G, namely, Grant Based (GB), Semi-Persistent Scheduling (SPS), and Grant Free (GF) scheduling. GB uses a handshaking procedure between the user and the gNode-B (gNB) [4]. SPS is used with periodic traffic [5]. In GF scheduling, the gNB reserves a certain BW for a user, or a group of users, to send their packets in an arrive-and-go manner without a handshaking procedure [6]. These techniques can be used interchangeably in the system. To fulfill each traffic requirement, an optimization problem should be formulated. The scheduling optimization problem is complex due to the variety of the traffic types and their targeted QoS.

GB scheduling is used for the eMBB traffic since they have moderate latency requirements. In contrast, GB cannot be used with URLLC since the handshaking associated latency will violate its latency requirement. In addition, SPS is not suitable since this traffic is sporadic and cannot be predicted. On the other hand, GF seems like a promising solution to the URLLC traffic since there is no handshaking required between the device and the gNB. In addition, resources can be allocated for more than one user. The main drawback of GF scheduling is its susceptibility to collision, which occurs when multiple nodes try to simultaneously access the same frequency resources. Collisions can be mitigated using HARQ techniques. There are three main HARQ techniques that are adopted in literature; reactive, k-repetitions, and proactive schemes [7]. The first two techniques are accepted by the 3GPP as the HARQ for GF scheduling [8]. The proactive scheme requires high processing power to transmit and receive at the same time, which is not available for most C-MTCDs.

The resource allocation problem for the uplink traffic is considered by several recent studies, however, limited research studies consider the mix of both types of traffic and how it will affect the system performance. The theoretical framework for the coexistence of different types of traffics in the uplink direction is discussed in [9] and [10]. A general mathematical analysis for the three main HARQ techniques is discussed in [7]. Some work considers only the URLLC requirements in the system as in [11]. It discusses the Random Access (RA) procedure in 5G communications and the URLLC traffic requirements for Factories of the Future (FOF). In [12], multi-users decoders are discussed to enable GF techniques. In [13], power boosting techniques in GF Uplink scheduling for URLLC are discussed. Different enhanced GF NOMA techniques are discussed in [14] to satisfy the URLLC requirements. Other work included the URLLC latency and reliability requirements in the uplink scheduling problem [15], [16]. In [15], a queuing model is developed for each service for both eMBB and URLLC. Another uplink optimization problem is discussed in [16]. The concepts of matching theory are applied to find sub-optimal solutions.

In this paper, we consider the problem of joint scheduling of the eMBB and URLLC traffic types. GB scheduling is adopted for eMBB scheduling while GF with k-repetitions HARQ is adopted for the URLLC traffic. Unlike the previous studies, we focus on the interaction between both types of traffic and formulate a resource allocation problem to account for the different QoS requirements. To the best of our knowledge, this is the first work done to model the probabilistic nature of the URLLC uplink traffic in a single cell. We derive an equation governing both the reliability and latency requirements for the URLLC devices based on the number of allocated RBs and the repetition factor. This equation will aid in choosing the optimum number of RBs and repetition factors to satisfy the reliability and latency requirements for URLLC devices. We analyze the effect of changing several system parameters on the outcome of the derived equation. In addition, we propose a novel scheduling algorithm that provides near-optimal results for the scheduling problem in real-time.

The rest of the paper is organized as follows. In Section II, the GF and GB subsystem models are formulated. In Section III, the optimization problem is introduced and the proposed solution procedure is discussed. Section IV analyzes the solution technique, provides a comparison between the scheduling techniques with the optimal scheduler, and discusses a solution to a complete operational scenario. Section V concludes the paper.

II. System Model

In this section, we develop both the GF and GB models for a single gNB in order to formulate the resource allocation optimization problem. Our system is composed of N_a URLLC devices, E eMBB users, and a single gNB. The GF URLLC devices share a common pool of resources as scheduled by the gNB at each Transmission Time Interval (TTI).[1] The frequency-time grid of each TTI is composed of N_f frequency slots and N_t short-TTIs (sTTIs). At the beginning of each TTI, the gNB broadcasts the minimum GF slots locations for all N_a users. The provided slots should, at least, fulfill their latency and reliability requirements. The k-repetitions is adopted as the HARQ scheme. In the k-repetitions, the c-MTCD sends k replicas of the packet one by one each sTTI and waits for feedback from gNB after transmission; the procedure is summarized in Figure 1. At the beginning of each TTI, the gNB receives the eMBB scheduling requests and then decides their RBs allocations to provide their rate requirements.

A. GF Model

The GF URLLC devices are permitted to send their packets in arrive and go manner; where at each time they have a packet to send, it is sent directly to gNB in the shared resource pool. It is assumed that the packet is sent upon arrival in the next sTTI and it is assumed that the transmission time is 1 sTTI as shown in Figure 1. In addition, the packet processing and feedback processing at gNB takes one sTTI each. The packet generation is modeled as a Bernoulli process with arrival probability p_a. Our channel is modeled as a flat fading Rayleigh channel with channel gain h. The receiver noise is modeled as additive white Gaussian noise with variance $\sigma^2 = N_0 B$. The packet is considered damaged when the Signal to Interference and Noise Ratio (SINR) is below the decoding threshold, $SINR \leq \gamma_{th}$. Since the gNB does not know the locations of the c-MTCDs, we define ρ as the full path-loss inversion power control that will compensate for the worst-case scenario path-loss and g_m is the m-th re-transmission power level where $g_m\rho$ is the targeted received power at gNB. We can define the maximum allowable re-transmissions before the latency constraint is violated as M. During the M re-transmissions, the c-MTCDs can re-transmit their packets, with the associated k replicas, without violating the maximum allowable latency, τ. The value of M can be calculated as

$$M = \left\lfloor (\tau - 1)/T^{RTT} \right\rfloor, \tag{1}$$

where $T^{RTT} = k + 3$ sTTI is the round trip transmission time for the k-repetitions scheme, the three sTTIs are added based on our transmission, processing, and feedback delays assumptions, and τ is the maximum latency, in sTTI units, for URLLC devices. Define the Probability of Delay Bound Violation (PDBV) as

$$P_F = P(T \geq \tau) = \begin{cases} 1 & M = 0 \\ 1 - \sum_{m=1}^{M} A_m p_m & M \geq 1 \end{cases}, \tag{2}$$

where A_m is the probability that the URLLC device is still active in the m-th round trip if the last $m-1$ re-transmissions failed. And p_m is the GF access success probability, as defined in [7]. It can be shown that

$$A_m = \begin{cases} 1 & m = 1 \\ 1 - \sum_{i=1}^{m-1} A_i p_i & m \geq 2 \end{cases}, \tag{3}$$

and

$$\begin{gathered} p_m = \sum_{n=0}^{N_a} \binom{N_a}{n} (A_m p_a/R)^n (1 - A_m p_a/R)^{N_a - n} \\ .\Theta[n,m,k](1 - \Theta[n,m,k])^n \end{gathered}, \tag{4}$$

where R is the number of allocated resources to URLLC devices and

$$\Theta[n,m,k] = 1 - \prod_{l=1}^{k} (1 - P(\mathrm{SINR}_l^m \geq \gamma_{th}), \tag{5}$$

where $\Theta[n,m,k]$ is the transmission success probability of the URLLC device given that the number of interfering URLLC devices equals n. SINR_l^m is the signal to noise and interference ratio of the l^{th} replica of the m^{th} re-transmission. It is assumed in our analysis that hard decoding is applied. If at least one of the k replicas is decoded then the transmission is considered a success. Equation (5) can be re-written using the Laplace transform for aggregate interference received at gNB as

$$\Theta[n,m,k] = \sum_{l=1}^{k} (-1)^{l+1} \binom{k}{l} \frac{\exp(-l\gamma_{th}\sigma^2/g_m\rho)}{(1+\gamma_{th})^{ln}}. \tag{6}$$

The complete proof of the derived equations is omitted due to space limitations. The equations developed in this section will be used to decide on the optimal number of resources, R, and the number of repetitions, k, to achieve the targeted latency and reliability requirements of URLLC devices with the least possible resources to maximize the eMBB QoS.

B. GB Model

The gNB needs to schedule the GB eMBB users along with giving enough resources to URLLC devices to meet their delay bounds. Since we aim to maximize the rate of eMBB users,

[1] A TTI is 1 msec as defined in the 3GPP standards [8].

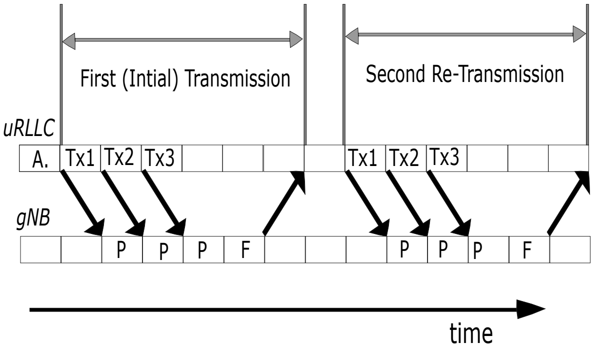

Figure 1: k-repetitions diagram with $k = 3$

we need to define the rate equations based on a scheduling parameter S^e_{ij}; the scheduling parameter, S^e_{ij}, is defined as

$$S^e_{ij} = \begin{cases} 1 & \text{eMBB user } e \text{ is allocated the } (i,j) \text{ RB} \\ 0 & \text{elsewhere.} \end{cases} \tag{7}$$

The eMBB user rate is given by

$$R_e = \sum_{i,j} S^e_{ij} B \log(1 + SNR^e_{ij}), \tag{8}$$

where $SNR^e_{ij} = \frac{|h_{ij,e}|^2 P_e}{N_0 B}$. The symbol $h_{ij,e}$ denote the channel coefficient and P_e is the transmission power of the e^{th} eMBB user. In the next section, the equations developed in this section are used to formulate the gNB resource scheduling optimization problem.

III. Problem Formulation and Proposed Algorithm

Based on the previous discussion, a good scheduler should provide the highest rates possible to eMBB users, while guaranteeing a minimum rate for each user. In addition, it should provide the minimum possible resources to satisfy the reliability and latency requirements of the URLLC devices. Failing to do this may result in under-utilizing the network resources and compromising the users' experience. Based on the system model described in the previous section and the developed equations, we can define our optimization problem that aims to maximize the eMBB users rate while satisfying the PDBV requirements of the URLLC devices. We can define our resource allocation optimization problem as

$$\max_{S^e_{ij}, R, k} \sum_e R_e \tag{9}$$

subject to

$$p(T \geq \tau) \leq \epsilon, \tag{9a}$$

$$S^e_{ij} \in \{0, 1\}, \qquad \forall i, j, e \tag{9b}$$

$$R_e \geq R^{min}_e \qquad \forall e \tag{9c}$$

$$\sum_e S^e_{ij} = 1, \qquad \forall i, j \tag{9d}$$

$$R \in \mathbb{N} \tag{9e}$$

$$S^e_{ij} = 0, \qquad \text{for } i \in \{i_1, i_2, ..., i_R\}, \forall j, e \tag{9f}$$

Equation (9) aims to maximize the overall rate of all eMBB users. Equation (9a) is the PDBV for URLLC devices with maximum allowable error ϵ. Equation (9c) is used to prevent the starvation of any of the scheduled eMBB users by guaranteeing a minimum rate R^{min}_e. Equation (9d) constraints the number of scheduled eMBB users on each RB to only one user. Equation (9e) ensures that the number of allocated resources to URLLC devices belongs to the set of Natural numbers. The non-multiplexing constraint between different types of traffic is defined in Equation (9f), where i_r is the number of allocated frequency resources to URLLC devices.

The optimization problem defined is a mixed-integer non-linear programming problem. As such, it cannot, in general, be solved using normal optimization methods. The problem is divided into two sub-problems as follows; satisfying the URLLC device's requirements with minimum possible resources and optimal scheduling for eMBB users. The separation done will not affect the optimality of the problem, since the PDBV is only affected by the number of the allocated URLLC slots, R.

In the first part, Equation (9a) is solved for a specific number of allocated slots to the URLLC devices, R, and repetitions, k. We will select the minimum R (and the corresponding k) to satisfy the URLLC devices latency requirements and this will, in turn, result in maximizing the eMBB users rates; this is because, from the eMBB users point of view, we have wasted the least amount of frequency resources to meet the URLLC devices latency constraints. Given the latency threshold, the number of maximum re-transmissions, M, is calculated. Using the iterative process, A_m and p_m are calculated for every possible value of the packet replication, m. If the inequality is satisfied, the scheduling step is executed. If not, the value of R is increased or the value of k is changed to fulfill the reliability requirements. In the next step, the positions for these resources are chosen and a scheduling policy of the remaining resources for all eMBB users is chosen in order to maximize the overall eMBB rate while satisfying the minimum rate requirements.

An optimal solution can be found using the grid search method by searching over all possible URLLC allocations for the optimal eMBB allocation that provides higher rates while satisfying the minimum rate requirements. It requires high processing power, which makes it unfeasible for operational environments. The best Channel Quality Indicator (best CQI) scheduling algorithm [17] is considered a greedy algorithm that aims to maximize the overall system throughput without considering the minimum rate requirements. Therefore, the users with bad channel conditions will starve for resources. In contrast, the Proportional Fair (PF) scheduling [17] maintains fairness among eMBB users as they are scheduled in a way that guarantees an almost equal rate to all eMBB users. However, it results in decreasing the overall throughput of the network. Another family of commonly used approaches for resource allocation in wireless networks is the Genetic Algorithm (GA) based approaches. The GA approach tries to solve the same problem using reproduction and mutations. The GA scheduler ensures that the starvation problem is solved by guaranteeing a minimum rate for each eMBB user, in addition to maximizing the overall system throughput. The GA-based approaches are considered to be search techniques; yet, they are, in general, faster than grid search-based approaches. However, in larger dimensions, they require high computation power.

A. Proposed Scheduling Technique

Due to the shortcomings of the previously discussed techniques, we propose a new technique to capture the best characteristics of both the best CQI and PF techniques. The first step is to split the resources equally between the eMBB users, by assigning $N_{ch} = \frac{N_f - R}{E}$ channels for each user, to ensure the minimum rate requirement is fulfilled. Then the best assignment of these resources is done to ensure the maximum overall rate is reached. The users are ordered randomly, to avoid extra processing, and for each iteration, each user is assigned the best, highest gain channel based on the remaining resources. Algorithm 1 shows the pseudo-code of our proposed scheduling algorithm.

Algorithm 1 Proposed Algorithm for the Channels Assignment problem

```
Require: $R$ and CSI
    Calculate $N_{ch}$, the number of channels assigned to each
    eMBB users to ensure minimum rate requirements
    for $j = 1$ to $N_{ch}$ do
        for $i = 1$ to $E$ do
            Choose the best $j^{th}$ channel for user $i$ using CSI
        end for
    end for
    Reserve the worst R channels for URLLC traffic
```

IV. Evaluation Results

In this section, we first analyze the developed PDBV equation. Then, we compare different scheduling techniques with the optimal scheduler. Finally, we analyze an operational scenario and compare different scheduling results. Unless stated differently, the system parameters are given in Table I.

Table I: system Parameters

Parameter	Value	Parameter	Value	Parameter	Value
BW	100 MHZ	N_a	150	N_t	10
N_f	100	ϵ	10^{-5}	τ	1.4 ms
γ_{th}	0.1	P_e	0.5 W	σ^2	-114 dbm
P_a	10^{-5}	g_m	m	R_e^{min}	12 Mbps

A. Analysis of PDBV for GF Traffic

In this section, we analyze the PDBV for URLLC devices, in Equation (9a), when varying different parameters, e.g., the repetition factor, k, the number of assigned frequency slots, R, the SINR threshold, γ_{th}, and the latency threshold, τ. The number of URLLC devices adopted, N_a, in this section, is st to be 50, and their latency threshold, τ, is set to be 1 ms. Figure 2 shows that at a low delay threshold increasing the value of k will negatively affect the system performance. In contrast, increasing the repetition factor, k, for c-MTCDS that have a higher latency threshold will decrease the PDBV and enhance the reliability. Figure 3 shows that increasing the latency threshold, τ, will not increase the PDBV. It is important to understand that sometimes increasing the latency will not affect the PDBV, because the maximum number of re-transmissions, M, is not changed. Finally, Figure 4 shows that as the SINR threshold increases, the decoding of the URLLC packets becomes difficult and the PDBV increases. In addition, the gap between one reserved frequency slot and two frequency slots, $R = 1$ and $R = 2$, curves shrinks as the SINR threshold, γ_{th}, increases. It should be noted that at medium SINR thresholds, some repetition values, $k > 1$, have better performance, but studying this aspect is out of our scope in this paper.

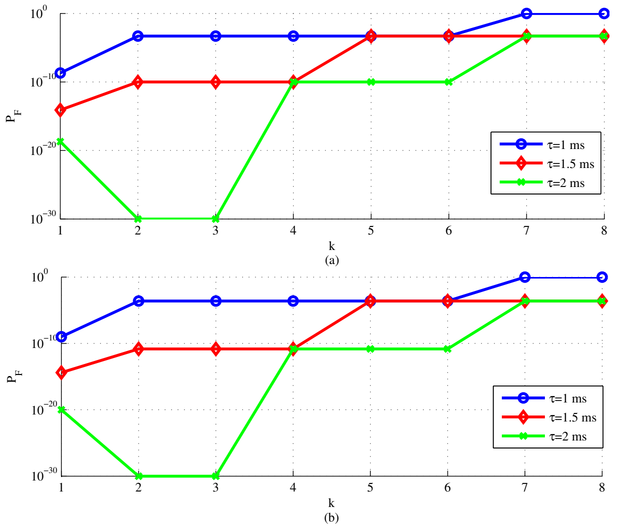

Figure 2: PDBV (P_F) vs the number of packet repetitions (k). (a) One RB ($R = 1$). (b) Two RBs ($R = 2$)

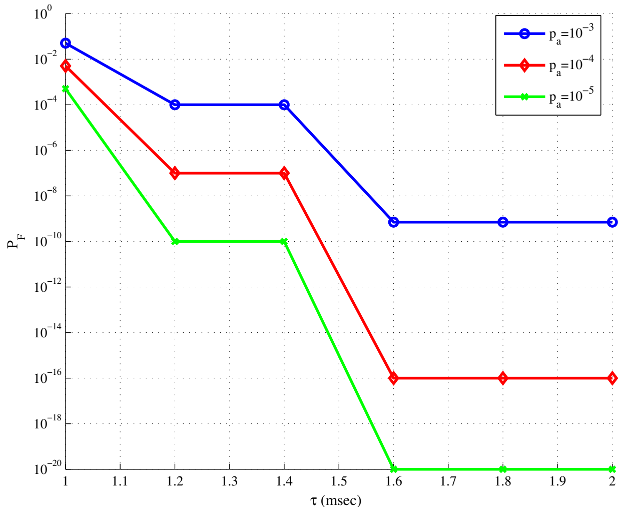

Figure 3: PDBV (P_F) vs the latency threshold (τ). The number of allocated RB (R)=1, and number of repetitions (k)=2

B. Optimal Scheduling for GB eMBB Traffic

In this section, we compare different scheduling algorithms with the optimal grid search technique. In this setup, we assume the number of frequency slots, N_f, to be equal to 6 and the minimum rate requirement for each eMBB user to be 2 Mbps with the rest of the parameters adopted from Table I. The goal of the scheduling algorithm, after knowing the number of frequency slots allocated for URLLC devices, is to choose the suitable channels that maximize the eMBB rate, while maintaining the minimum rate requirements for eMBB users. Figure 5 shows a comparison between the aforementioned algorithms, with a 95% Confidence Interval (CI). Figure 5a

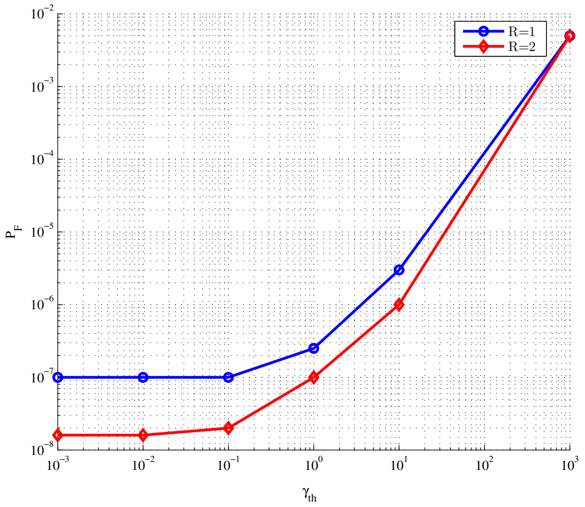

Figure 4: PDBV (P_F) vs the SINR threshold (γ_{th}). The packet arrival probability (p_a)=10^{-4}, and no repetitions, (k)=1

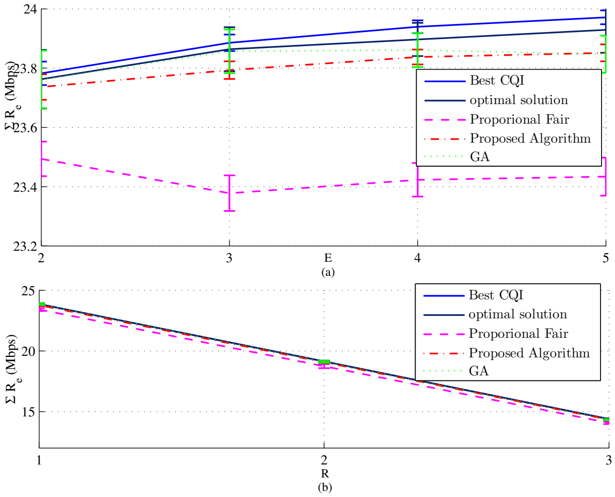

Figure 5: Comparison among the different Scheduling Techniques. (a) Fixing the number of URLLC allocated RBs, $R = 1$. (b) Fixing the number of eMBB users in the system, $E = 3$.

shows the accumulated eMBB rate when varying the number of eMBB users, E, and reserving one frequency slot for URLLC traffic, $R = 1$. In Figure 5b, the number of allocated URLLC frequencies, R, is changed, while maintaining the number of eMBB users fixed, $E = 3$. The Best CQI algorithm shows a higher accumulative rate than the optimal grid search since it ignores the minimum rate constraint for eMBB users. The GA achieves near-optimal performance in the case of a small search space as it can efficiently search for the optimal solution in this case. The proposed algorithm performs slightly lower than the GA to maintain the fairness condition among eMBB users. The PF performs the least due to the strict fairness condition.

In the next section, an operational scenario is discussed and it will be shown that our proposed algorithm performs better than the GA scheduler, the Best CQI, and the PF schedulers.

C. Operational Scenario Results

In this section, a full operational scenario is discussed. The system parameters are as summarized in Table I.

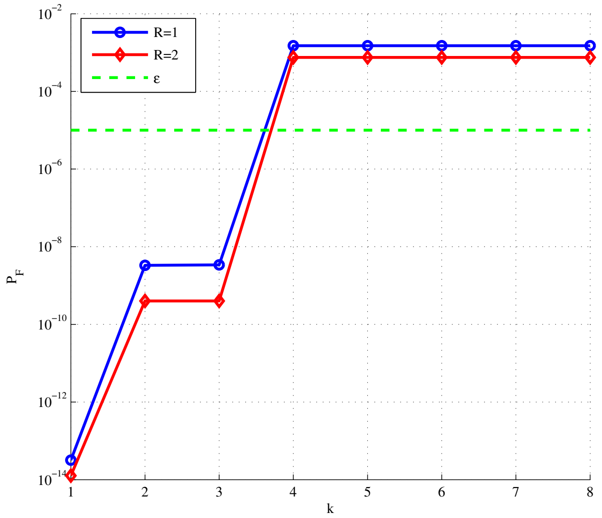

Figure 6: PDBV (P_F) vs the number of packet repetitions (k)

First, PDBV is calculated for different repetition factors, k, and different numbers of allocated frequencies, R. Figure 6 shows the results for $k = 1, 2, 3$ and $R = 1$ or $R = 2$. It should be noted that the reliability threshold accepted by the 3GPP for URLLC devices is 10^{-5} [8]; it is shown in Figure 6 to indicate the combinations of R and k that satisfy this reliability requirement.

Next, Best CQI, PF, GA, and our proposed algorithm are used for the scheduling step. In addition, to show the drawbacks of Best CQI, an error percentage is calculated in each case where the results are taken by averaging several simulation runs. In addition, a CI of 95% is calculated for each case. Figure 7a shows the accumulated eMBB rate when varying the number of eMBB users, E, and reserving one frequency slot for URLLC traffic, $R = 1$. While Figure 8a shows the accumulated eMBB rate when varying the number of allocated URLLC frequencies, R, for a fixed number of eMBB users, $E = 15$.

As shown in Figures 7 and 8, the Best CQI algorithm results in the highest sum data rate. However, the algorithm violates the minimum rate requirements as shown in Figures 7b and 8b, and the number of violations increases as the number of eMBB users increases. Our proposed algorithm comes second in terms of the sum data rate, with all the requirements satisfied. The GA approach produces results that are lower than our approach, due to the high dimension of the problem in this case of the operational scenario. Moreover, our proposed approach is more efficient and requires less computational resources as compared to the GA based approach. The PF algorithm remains the least one in terms of the overall achieved sum data rate due to its strict fairness condition.

V. Conclusions

In this paper, we address the problem of mixed eMBB and URLLC traffic in 5G networks. For this purpose, the k-repetitions HARQ GF scheduling technique is discussed and the probability of delay bound violation for any number of URLLC devices is derived for a single cell. The results show the effect of changing several system parameters on

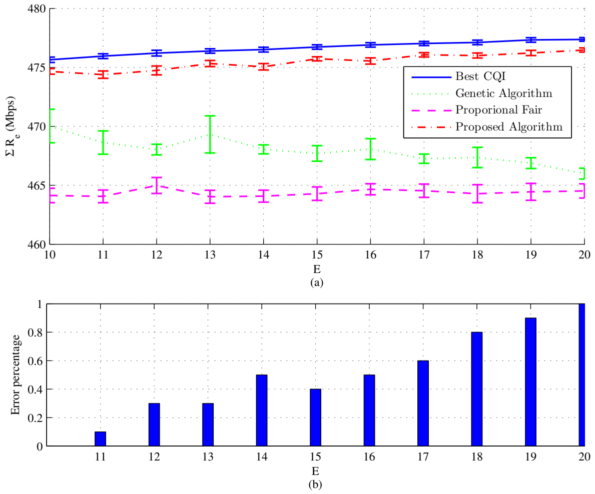

Figure 7: (a) Comparison among the different scheduling algorithms. (b) Error percentage in satisfying the minimum rate requirements.

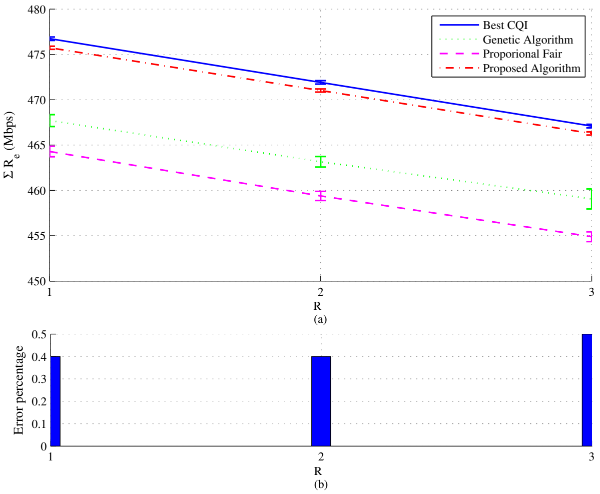

Figure 8: (a) Comparison among the GA, PF, Best CQI, and the proposed algorithm. (b) Error percentage in satisfying the minimum rate requirements.

the probability of the delay bound violation. An optimization problem for uplink scheduling is formulated with the aim to maximize the eMBB rate while satisfying the reliability and delay requirements of URLLC devices. In addition, a minimum rate guarantee for each eMBB user is maintained. The formulated problem is a mixed-integer non-linear optimization problem. The techniques that can produce optimal results for this family of problems require high computational power and time. This is not suitable for real-time applications. Therefore, a simple scheduling technique is proposed, which benefits from both the qualities of the PF algorithm and the Best CQI algorithm. Evaluation results show that the proposed scheme can *efficiently* result in a near-optimal performance that is better than other more complex algorithms under different operating conditions.

References

[1] M. Series, “Minimum requirements related to technical performance for imt-2020 radio interface (s),” *Report*, pp. 2410–0, 2017.

[2] G. T. 38.913, “Study on scenarios and requirements for next generation access technologies,” 2016.

[3] M. Series, “Imt vision–framework and overall objectives of the future development of imt for 2020 and beyond,” *Recommendation ITU*, vol. 2083, 2015.

[4] G. T. . V14.2.0, “Evolved universal terrestrial radio access (e-utra); physical layer procedures,” 2017.

[5] D. Jiang, H. Wang, E. Malkamaki, and E. Tuomaala, “Principle and performance of semi-persistent scheduling for voip in lte system,” in *2007 WICOM*, 2007, pp. 2861–2864.

[6] T. Fehrenbach, R. Datta, B. Göktepe, T. Wirth, and C. Hellge, “Urllc services in 5g low latency enhancements for lte,” in *2018 IEEE 88th Vehicular Technology Conference (VTC-Fall)*, 2018, pp. 1–6.

[7] Y. Liu, Y. Deng, M. Elkashlan, A. Nallanathan, and G. K. Karagiannidis, “Analyzing grant-free access for urllc service,” *IEEE Journal on Selected Areas in Communications*, pp. 1–1, 2020.

[8] 3GPP, “NR; NR and NG-RAN Overall description; Stage-2,” 3GPP, Technical Specification (TS) 38.300, 03 2019, version 15.5.0. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3191

[9] P. Popovski, K. F. Trillingsgaard, O. Simeone, and G. Durisi, “5g wireless network slicing for embb, urllc, and mmtc: A communication-theoretic view,” *IEEE Access*, vol. 6, pp. 55 765–55 779, 2018.

[10] R. Kassab, O. Simeone, and P. Popovski, “Coexistence of urllc and embb services in the c-ran uplink: An information-theoretic study,” in *2018 IEEE GLOBECOM*, 2018, pp. 1–6.

[11] J. Thota and A. Aijaz, “On performance evaluation of random access enhancements for 5g urllc,” in *2019 IEEE WCNC*, 2019, pp. 1–7.

[12] C. Wang, Y. Chen, Y. Wu, and L. Zhang, “Performance evaluation of grant-free transmission for uplink urllc services,” in *2017 IEEE 85th VTC*, 2017, pp. 1–6.

[13] R. Abreu, T. Jacobsen, G. Berardinelli, K. Pedersen, I. Z. Kovács, and P. Mogensen, “Power control optimization for uplink grant-free urllc,” in *2018 IEEE WCNC*, 2018, pp. 1–6.

[14] N. H. Mahmood, R. Abreu, R. Böhnke, M. Schubert, G. Berardinelli, and T. H. Jacobsen, “Uplink grant-free access solutions for urllc services in 5g new radio,” in *2019 16th ISWCS*, 2019, pp. 607–612.

[15] L. Feng, Y. Zi, W. Li, F. Zhou, P. Yu, and M. Kadoch, “Dynamic resource allocation with ran slicing and scheduling for urllc and embb hybrid services,” *IEEE Access*, vol. 8, pp. 34 538–34 551, 2020.

[16] M. Y. Abdelsadek, Y. Gadallah, and M. H. Ahmed, “Matching-based resource allocation for critical mtc in massive mimo lte networks,” *IEEE Access*, vol. 7, pp. 127 141–127 153, 2019.

[17] M. H. Habaebi, J. Chebil, A. Al-Sakkaf, and T. Dahawi, “Comparison between scheduling techniques in long term evolution,” *IIUM Engineering Journal*, vol. 14, no. 1, 2013.