Wireless Networks
https://doi.org/10.1007/s11276-023-03328-2

ORIGINAL PAPER

Resource allocation scheme for eMBB and uRLLC coexistence in 6G networks

Muhammed Al-Ali[1] · Elias Yaacoub[1]

Accepted: 21 March 2023

Abstract

5G technology is intended to support three promising services with heterogeneous requirements: Ultra-Reliable and Low Latency Communication (uRLLC), enhanced Mobile Broadband (eMBB), and massive Machine Type Communication (mMTC). 6G is required to support even more challenging scenarios, including the presence of a large number of uRLLC devices, under the massive uRLLC (mURLLC) use case scenario. The presence of these services on the same network creates a challenging task of resource allocation to meet their diverse requirements. Given the critical nature of uRLLC applications, uRLLC traffic will always have the highest priority which causes a negative impact on the performance of other services. In this paper, the problem of uRLLC/eMBB resource allocation is investigated. An optimal resource allocation scheme is proposed with two scenarios including a guaranteed fairness level and minimum data rate among eMBB users. In addition, a knapsack-inspired punctured resource allocation algorithm is proposed where the users' channel qualities of both services are considered at each time slot leading to the most suitable Resource Block (RB) selection for puncturing in a way that minimizes the negative impact on eMBB performance. The proposed solution was compared with three puncturing baseline reference algorithms and the performance was evaluated in terms of eMBB Sum throughput and Fairness level. The simulation results show that the proposed algorithm outperforms the above-mentioned reference algorithms in all evaluation metrics and is proved to be comparable to the optimal solution given its low complexity.

Keywords 5 G/6 G · uRLLC · eMBB · Resource allocation

1 Introduction

The massive technological development in electronic devices facilitated the emergence of new applications (e.g., Artificial intelligence (AI), Big Data analysis, the Internet of everything, Virtual Reality (VR), etc.) having a ubiquitous influence on people's lives. Nonetheless, these applications produce a huge amount of data traffic in addition to requiring continuous connectivity, raising one of the most challenging tasks for today's cellular communication technologies to overcome. It is expected for Smart Phones, Tablets, Routers and Mobile PCs combined data traffic to reach 169 exabytes/month by the end of 2027 compared to 18 exabytes/month in 2022 [1], which certainly frames the technical objectives of the future cellular system. In addition, the supported applications like Virtual reality (VR), Remote surgery, Intelligent Transportation Systems, High Voltage Electricity distribution, and Industrial Control have different requirements. It is clear that handling these heterogeneous requirements is a challenging task and considered one of the reasons the International Telecommunication Union (ITU) classified the services 5 G is envisioned to support into different categories [2, 3]. The first one is enhanced Mobile Broadband (eMBB) which aims to provide high data rates, high user mobility, and better connectivity which is essential for human-centric applications [3, 4]. The second category is massive Machine Type Communications (mMTC), designed to provide efficient connectivity to a massive number of devices and it is considered one of the

Muhammed Al-Ali and Elias Yaacoub have contributed equally to this work.

✉ Elias Yaacoub
elias.yaacoub@gmail.com

Muhammed Al-Ali
muhd.s.yaseen@gmail.com

[1] Department of Computer Science and Engineering, Qatar University, Doha, Qatar

Published online: 16 April 2023

main enablers of the Internet of Things (IoT) [5]. Lastly, ultra-Reliable and Low Latency Communication (uRLLC), which is specifically designed for mission-critical applications targeting 99.999% reliability and as low as 1 ms latency [6]. Moreover, 6 G is envisioned to support massive uRLLC (mURLLC) combing the requirements of mMTC and uRLLC use cases realizing that IoT devices are expected to reach a massive number of 25 billion in 2025 [7, 8]. This illustrates the size of traffic the BS might encounter that is likely to belong to mission-critical applications requiring high priority over others. While classifying these services into different classes helps in identifying the applications by which these services are used and thus assigning suitable priority levels for each in order to support all applications efficiently, this created a new obstacle towards achieving the best operational performance. The coexistence of these services with their heterogeneous requirements within the same network infrastructure creates a challenging resource allocation problem given that the operators are tied by a finite bandwidth (BW) and limited operational cost budget to satisfy their Quality of Service (QoS) requirements. Thus, providing the QoS requirements of uRLLC users will automatically reduce the resources available for the existing eMBB users. What complicates the situation further is the fact that uRLLC has a stochastic nature in which it has unexpected arrival at the BS that needs to consider the size of the incoming traffic, the availability of the resources, and the unstable radio channel conditions while being forced to serve the uRLLC momentarily given its strict latency and reliability requirements. The 3rd Generation Partnership Project (3GPP) proposed two scheduling approaches to handle the uRLLC traffic. The first approach is known as reservation-based scheduling while the other one is known as instant scheduling (Preemptive/Puncturing scheduling) [9–11]. The first approach uses a uRLLC reservation-based frame to handle any unexpected traffic. It can either use static or dynamic resource reservation. Static reservation method tends to send the frame structure that holds the transmission configurations (e.g., adapted Numerology) in an intermittent fashion. Unlike the static reservation method, in dynamic reservation, the frame structure is sent frequently to the UE. This approach causes a control signaling overhead and the resources reserved for the uRLLC might be wasted in the case where there are no incoming uRLLC data. The second approach (known as Instant scheduling) aims to serve any incoming uRLLC traffic instantly using short Transmission Time Intervals (TTIs) of 2,4,7 OFDM symbols (mini slot-based scheduling) [12, 13]. While this approach might cause an interruption to ongoing transmissions of other applications and might cause huge performance degradation of other services, it is still considered a more efficient approach as it can be relied on to support the strict latency requirements of uRLLC.

Consequently, it is important to investigate joint optimization of resource alloation for both eMBB and uRLLC, in order to satisfy the stringent requirements of uRLLC without impacting (or while minimizing the impact on) eMBB users. Such a joint optimization gains additional importance in a beyong 5 G/6 G scenario, where an increased number of uRLLC users in an mURLLC scenario further impacts eMBB users. In our previous work [14], a basic attempt was made to investigate joint resource allocation for eMBB and uRLLC, where the problem was formulated and a single suboptimal algorithm was proposed. In this paper, we provide significant enhancements by (i) considering both the problem formulation and deriving the optimal solution, (ii) proposing four suboptimal algorithms and comparing their performance, (iii) implementing additional fairness criteria, and (iv) generating an extensive set of simulation results.

Hence, in this paper, both optimal and sub-optimal approaches have been adopted aiming to find the most suitable RB at every time slot for puncturing according to different criteria. The main common constraint in both approaches is satisfying the requirements of the existing uRLLC traffic at each time slot. The main contributions of this work are summarized as follows:

1. Formulating the problem of resource allocation as an optimization problem aiming to maximize the sum throughput of eMBB users in the presence of uRLLC traffic with different intensities and outage probabilities,
2. Deriving the solution of the optimization problem under two constraints where a pre-defined fairness level or a guaranteed minimum data rate among eMBB users is enforced in each scenario,
3. Proposing different practical sub-optimal resource allocation algorithms that perform efficient resource allocation in a joint eMBB-uRLLC scenario, while taking different constraints into consideration, and
4. Comparing the performance of the proposed algorithms in various scenarios and analyzing their performance tradeoffs.

The rest of the paper is structured as follows. In Sect. 2, we review some of the related works in the literature. In Sect. 3, we introduce the system model and formulate the problem. In Sect. 4 we provide and discuss the numerical simulation results. Finally, we present the conclusion in Sect. 5.

2 Related work

Several references have investigated resource allocation in 5 G and beyond networks. For example, in [15], joint scheduling of guaranteed bit rate (GBR) and non-GBR services in 5 G was studied, although eMBB users only were considered for both cases (GBR and non-GBR). Resource allocation for 5 G and beyond was studied in [16], where heterogeneous networks with macrocells, small cells, and femtocells were considered. However, uRLLC was not considered except in the discussion of network slicing, but joint eMBB-uRLLC scheduling was not investigated. Similarly, uRLLC slices were mentioned in [17, 18], where the focus is on 5 G vehicular scenarios (V2X), whereas [19] investigated 5 G IoT scenarios.

In this section, we discuss the most relevant related works that highlight the problem of resource allocation of eMBB and uRLLC traffic. Different techniques are studied in this section that involves both instant and reservation-based scheduling. In [20], the authors propose an online joint scheduling framework algorithm of eMBB and uRLLC, formalizing and solving the problem of resource puncturing on eMBB traffic. The authors used different models to tackle this problem. The linear model is used when the degradation in eMBB data rates is directly proportional to the amount of punctured resources in which an optimal resource scheduling algorithm is introduced. The scheduler targets the stochastic nature of uRLLC traffic and aims to place it in a uniform random fashion in each slot while scheduling the eMBB UE via an iterative greedy method that considers the expected degradation in eMBB data rates. The Convex model is used when the uRLLC traffic can be modeled as a convex function. The decomposition of this model is not as efficient as the linear model making an optimal allocation more difficult. This led the authors to adopt a simpler uRLLC traffic placement model which is fixed across the whole time slot (across all mini-slots). In [21], the authors propose a downlink scheduling algorithm that aims to satisfy a minimum achievable eMBB data rate with an optimal resource allocation for uRLLC traffic. They address eMBB and uRLLC users with pending retransmissions with uRLLC having the highest priority in order to satisfy the reliability constraint. Maximizing the minimum eMBB data rate is based on two preferences in which the first one is the expected eMBB data rate till time slot t. The second preference is based on the uRLLC placement strategy which is derived according to historical uRLLC latency and reliability demands. The resource allocation decision is based on these two metrics and the results show a noticeable improvement in this approach over random resource allocation schemes. In [22], the authors formulated the resource allocation problem in terms of the eMBB data rate and uRLLC interrupt probability requirement. The proposed approach includes a resource block allocation scheme that satisfies the reliability requirements of uRLLC. This reliability is evaluated by measuring the transmission power of uRLLC users and the users' outage probability. The proposed scheme is based on the allocation of RBs where uRLLC users experience the best channel conditions and the simulation results indicated that both maximizing the eMBB data rate and satisfying the reliability requirements of uRLLC can be achieved using the proposed algorithm. In [23], resource allocation was formulated as an optimization problem. The authors' approach was based on superposition and puncturing schemes governed by the preference profiles of eMBB and uRLLC users. The preference profile is based on the ability of uRLLC users to tolerate the degradation of their QoS. The solution includes the classification of uRLLC users depending on their geographical location which helps in predicting the willingness of uRLLC users to opt for superposition instead of resource puncturing in a contract-based framework. In [24], the authors adopted the network slicing approach in which the resource allocation problem was formulated as a risk-sensitive form that aims to enhance the reliability of eMBB and uRLLC traffic. A deep reinforcement learning approach has been adopted for maximizing the average data rate of eMBB UE and minimization of eMBB data rate variance. The results indicated that the proposed work could satisfy the requirements of uRLLC while preserving the desired reliability level of eMBB users. In [25], the authors proposed a non-orthogonal multiple access (NOMA) based solution for the problem of resource allocation of eMBB and uRLLC. The solution depends on matching theory by finding the optimal pairs of users upon performing superposition, in order to satisfy the QoS requirements of uRLLC and maintain fairness among eMBB UE. Numerical results showed that the authors proposed work can provide a high minimum expected achieved data rate (MEAR) for eMBB UEs while preserving fairness using different 5 G NR numerologies. In [26], the authors proposed dynamic joint scheduling for eMBB and uRLLC traffic. The uRLLC latency requirement has been satisfied based on a queuing methodology, evaluated in terms of outage probability and throughput. The eMBB data rate was also maximized by deriving and solving an outage-constrained stochastic optimization problem where resource puncturing is adopted in the solution. The simulation results indicated that the proposed solution outperforms a non-queuing approach. In [27], the authors proposed a superposition-based approach in which one-to-one pairing of eMBB and uRLLC users is adopted aiming to overcome the high complexity of the original optimization problem. The solution has been evaluated in terms of eMBB data rate loss, uRLLC packet segmentation loss, and admission rate taking into consideration different loss thresholds for each case. The main pairing criteria is avoiding segmentation where a single uRLLC packet can only be paired with one eMBB user. The optimal pairing is based on power and resource allocation that minimizes eMBB data rate loss. In [28], the authors formulated an integer programming problem and solved it using two methods, convex relaxation, and a fairness-aware greedy algorithm. The resource allocation for uRLLC depends on the decisions of the eMBB scheduler. The objective is to select a suitable user satisfying the latency and reliability requirements represented by an end-to-end delay budget and probability of failure. The greedy approach is based on calculating the proportional fairness weight that depends on the previously achieved eMBB data rate and the priority is given to the RBs requested by the uRLLC user. In [29], the authors reformulated the problem of maximizing eMBB throughput into a conflict minimization problem between eMBB and uRLLC services. The solution leverages the free selection of multi-numerology offered in 5 G NR and the conflict minimization is considered a case of bin packing optimization. The eMBB throughput maximization is linked to the minimization of aggregated conflict upon the selection of each RB for uRLLC placement. The results show that high resource efficiency can be achieved using the proposed approach with linear complexity. In [30], the authors proposed a hybrid approach that maximizes the average throughput of eMBB users and the uRLLC admission rate. Their approach is based on superposition and puncturing for eMBB and uRLLC downlink transmission where they derive a sub-optimal solution to the problem using sequential convex programming. Maximizing the average data rate is used as a method to preserve fairness among eMBB users aiming to provide better spectral efficiency and a stationary eMBB QoS.

The main motivation behind this work is the fact that a lot of the reviewed papers in the literature are based on some assumptions that might limit their practicality in the case of uRLLC, especially when it deals with mission-critical applications. Some of the papers assume a higher delay budget for the uRLLC traffic or a certain level of processing power. Reinforcement learning for instance that is based on benefiting from previous experience when making decisions (learning via trial and error) might not necessarily work with uRLLC, e.g., when considering remote surgery that cannot tolerate even a small margin of error. Contract-based approaches between eMBB and uRLLC users do not consider the amount of uncertainty the BS has to deal with when assigning resources. This includes the uRLLC traffic density, arrival rate, QoS requirements, and channel quality of the uRLLC users. Moreover, Superposition based approaches, which are based on the idea of simultaneous transmissions by the BS to both eMBB and uRLLC users, are also found in the literature. Although this approach lowers the impact on eMBB users, it assumes the ability of the uRLLC end device to perform Successive Interference Cancelation (SIC). IoT is one example of devices with low processing power where SIC might be difficult to implement. Thus, this work addresses these limitations in the literature by adopting the resource puncturing approach. Four different algorithms are investigated: the first one achieves uRLLC requirements regardless of their impact on eMBB users. The second algorithm focuses on protecting the cell-edge eMBB users suffering from low data rates, whereas the third one aims to maximize the eMBB sum-throughput while achieving uRLLC requirements. Finally, the main novelty is in the fourth algorithm that strikes a delicate balance between maximizing eMBB sum-throughput and maintaining fairness among eMBB UEs, while meeting the QoS requirements of uRLLC UEs.

2.1 Heuristic scheduling algorithms

One of the key features of the current generation of communication networks is the Radio Resource Management (RRM) techniques that are utilized to improve system performance. One of the important parameters to achieve this desired improvement are Packet scheduling algorithms which play an important role in allocating resources represented by frequency and time to the connected users. These algorithms consider channel quality condition and the QoS requirements of these users when making the resource allocation decision aiming to provide an optimal tradeoff between system throughput, spectral efficiency, and fairness. These algorithms work at the base station and are responsible for allocating fractions of the spectrum to the connected users.

2.1.1 Proportional fairness in time and frequency (PFTF)

The main objective of this scheduler is to balance between maximizing the data rate and fairness among the users. It is considered a channel-aware and QoS-unaware scheduler. The utility function of this scheduler is written below [31, 32]:

$$U = \max_{i} \left[\frac{r_i(k,t)}{R_i(t) + \sum_{j=1}^{m} r_i(j,t)} \right] \tag{1}$$

where $\sum_{j=1}^{m} r_i(j,t)$ is the total data rate of user i along all the resource blocks $j = 1$ to m which are allocated to it at TTI t. The notation $r_i(k,t)$ is used to describe the instantaneous throughput of user i at time t (The amount of data the BS can transmit to user i at time t using RB k) which is directly dependent on the channel condition of user i. Equation (2) expresses the well-known Shannon capacity which is the data rate at which the data can be transmitted reliably (i.e., small error probability).

$$r_i(k,t) = BW_k \log_2[1 + SNR_i(k,t)] \tag{2}$$

$$r_i(t) = \sum_{j=1}^{m} r_i(j,t) \tag{3}$$

where BW is the available bandwidth allocated to RB k and $SNR_i(k,t)$ is the achieved Signal to Noise ratio of user i at RB k at time slot t assuming that user i is experiencing equal SNR in all subcarriers of RB k. $r_i(t)$ is the total data rate of user i over all its allocated RBs at time slot t. $R_i(t)$ is the average throughput of user i during a fixed size time window tc [33] as shown in Eq. (4).

$$R_i(t) = \left(1 - \frac{1}{tc}\right) * R_i(t-1) + \frac{1}{tc} * r_i(t) \tag{4}$$

2.1.2 Maximum-largest weighted delay first (M-LWDF)

This scheduler is designed to serve real-time users, and unlike PFTF, it considers the QoS requirements of these users including their Delay budget D_{QoS} and the lifetime expiration of their packets (Packet Loss Rate) PLR_{QoS} [34]. Its utility function is written below:

$$U = \max_i \left[Q_i * D_i(t-1) \frac{r_i(k,t)}{R_i(t)} \right] \tag{5}$$

$$Q_i = \frac{-\log(D_{QoS})}{PLR_{QoS}} \tag{6}$$

where Q_i is the parameter that considers the QoS requirements of user i and parameter $D_i(t-1)$ is the Head of the Line packet delay addressed to user i. In addition to those two algorithms, we considered the BestCQI allocation algorithm which aims to provide resources for the users with the highest Channel Quality Indicator (CQI).

2.2 Frame structure

6 G is expected to inherit one of the unique features of 5 G which is the flexible frame structure it offers. The frame structure is a grid of time and frequency in which the frequency domain is divided into a number of Resource Blocks (RBs) depending on the available Bandwidth. Each resource block includes 12 subcarriers. Moreover, different numerologies are supported where each numerology has a different value of subcarrier spacing (SCS). SCS equals to $15 * 2^M$ KHz (M can take a value between 0 and 4) and ranges between 15 and 240 kHz. Higher Subcarrier spacing values are used for higher frequencies in order to reduce the Inter-Carrier Interference (ICI). The time domain consists of subframes where each subframe might contain one or more time slots. The duration of each subframe is 1 ms while the duration of the time slots is scalable. The scalability of the time slot duration is subject to the size of the subframe and size of the SCS where these slots must not cross the boundary of the subframe and can range from 0.125 to 1 ms. The time slot usually consists of 14 OFDM symbols and the Cyclic prefix (CP) is used with different lengths (depending on the SCS) in order to mitigate the effect of Inter-Symbol Interference (ISI). Another important feature is the scalable TTI as the number of OFDM symbols per TTI can vary according to the network preference. This feature enables the scheduling of UE on Slot (14 OFDM symbols) and Mini-Slot (1–13 OFDM symbols) basis. TTI length can be adjusted by either reducing the number of OFDM symbols per TTI or by increasing the SCS and thus reducing the OFDM symbol duration. For instance, if the TTI is 0.125 ms, the UE can be scheduled in a slot-based fashion with an SCS of 120 kHz or they can be scheduled in a Mini-Slot based fashion by using an SCS of 15 kHz and mini-slot size of 3 OFDM symbols [6, 11–13]. Mini-Slot-based scheduling plays a crucial role in enabling uRLLC as the short TTI means a shorter processing time in addition to avoiding unnecessary delay, waiting to the next time slot for transmission.

3 System model and problem formulation

The system model is shown in Fig. 1. This figure provides an overview of the joint eMBB/uRLLC resource allocation scenario with puncturing, where the resources allocated to two eMBB UEs are punctured in order to serve a uRLLC UE that joins the network. In this section, we introduce the ideas of optimal and sub-optimal resource allocation, highlighting the qualities of each solution and explaining the reasons why one can be preferred over the other under specific scenarios.

3.1 Optimal resource allocation of eMBB/uRLLC traffic

In this part, we address the resource allocation problem of uRLLC and eMBB while considering the optimal solution to the problem. The aim is to highlight the importance of considering stochastic optimization and how the resource allocation problem is dealt with under uncertainty, which is caused by the mobility of the users, continuously changing propagation environments in addition to the availability of the resources themselves. We use a problem formulation adapted from [35–37] with an added constraint that

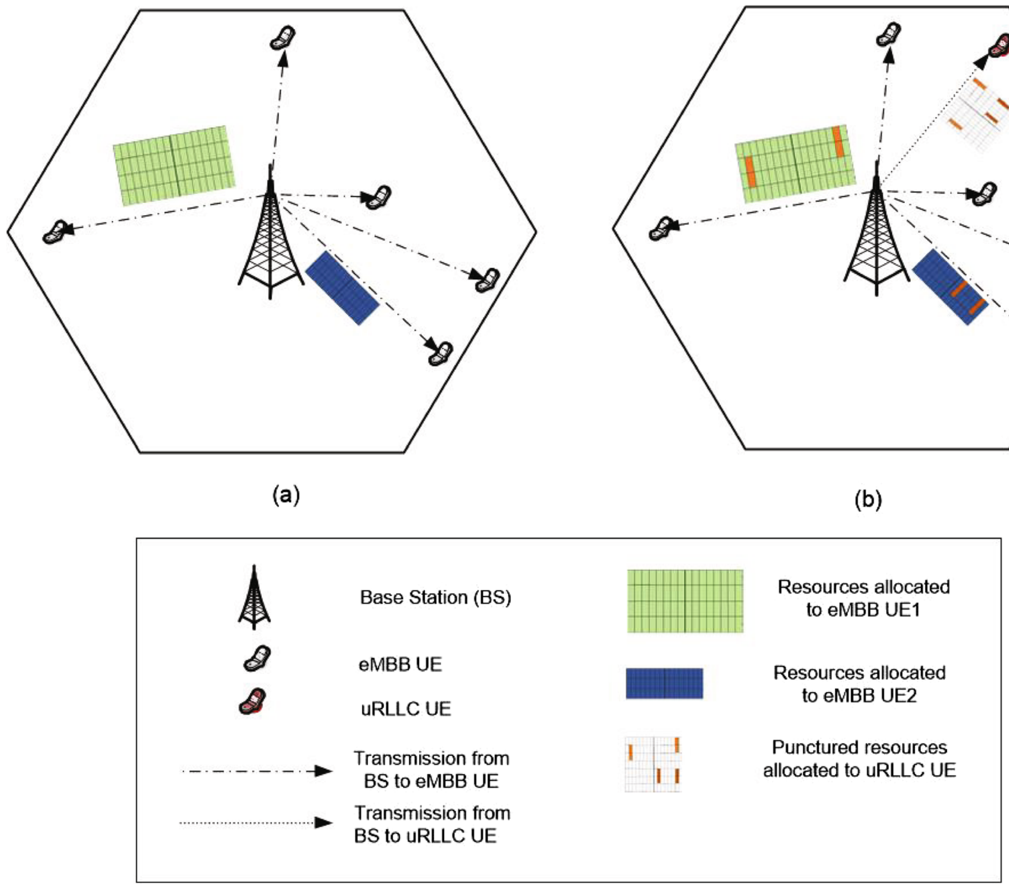

(a)
(b)
Base Station (BS)
eMBB UE
uRLLC UE
Transmission from BS to eMBB UE
Transmission from BS to uRLLC UE
Resources allocated to eMBB UE1
Resources allocated to eMBB UE2
Punctured resources allocated to uRLLC UE

Fig. 1 System model: **a** eMBB UEs only; **b** eMBB UEs with one uRLLC UE; punctured resources from two eMBB UEs are shown (to serve the URLLC UE)

guarantees a minimum data rate among eMBB UEs. The optimal allocation problem takes into consideration eMBB and uRLLC UEs in which the optimization variables are the fractions of *BW* each service acquires from the BS. The uRLLC payload is treated as a random variable derived from a random distribution with different rate parameters which form a chance (probability) constraint optimization problem. The probability constraint found in the optimization problem representing the uRLLC payload can be transformed into a deterministic form using the Cumulative distribution function (CDF) that corresponds to the random distribution used to represent the uRLLC flow (i.e., calculating the probability that the uRLLC payload will take a value equal or less than a specific size). We considered the desired outage probability, which uRLLC must not exceed, in the formulation. The problem formulation of our proposed work is shown below (we drop the time variable *t* to avoid overcrowding the equations, whenever no confusion can occur):

$$\max_{j} \sum_{i=1}^{n} (B_i - B_{iu}(j)) \log_2(1 + SNR_i) \tag{7}$$

s.t.

$$P\left[\sum_{j=1}^{m} B_{iu}(j) \log_2(1 + SNR_u(j)) < L_u\right] \leq \gamma \tag{7a}$$

$$\sum_{j=1}^{m} B_{iu}(j) \leq BW \tag{7b}$$

$$(B_i - B_{iu}(j)) \log_2(1 + SNR_i) \geq \boldsymbol{r} \quad \forall \mathrm{i} \in E \tag{7c}$$

where B_i and B_{iu} represent the amount of resources allocated and punctured to/from eMBB user i. SNR_i is the Signal to Noise ratio of eMBB user i. n is the number of eMBB UEs and m is the total number of mini-slots assigned to uRLLC UE. L_u is the size of uRLLC payload and γ represents the required reliability level of uRLLC traffic or it can be described as the confidence level for uRLLC users in which their data is transmitted within their latency budget (i.e., 1 ms). BW is the available bandwidth. The term E represents the set of eMBB users with RBs allocated to them at time slot t and $\boldsymbol{r}$ is the minimum data rate, pre-specified for each of those eMBB UEs. Constraint (7c) ensures a minimum data rate of $\boldsymbol{r}$ for each eMBB UE at time slot t.

In this approach, we are assuming that the bandwidth is fully used by the eMBB UEs and a uRLLC traffic arrives at each time slot with different payload sizes. The puncturing of resources is based on the size of the mini-slots or the TTI of the uRLLC while the initial allocation of the resources is a slot based with a length of 1 ms. A number of error probability thresholds were considered for the sake of comparing our work to [36].

It is important to mention that providing a reasonable fairness level among eMBB users cannot always be feasible because of the variation in these users' channel conditions. In other words, eMBB users with bad channel conditions might not be able to achieve a data rate close to those with better channel conditions, and forcing a certain fairness level on the optimization algorithm would lead to serious degradation in the sum throughput of the eMBB users in addition to the spectral efficiency. This is caused by forcing the optimizer to lower the data rates (i.e., provide fewer resources) to eMBB users with good channel condition in order to satisfy the desired fairness level which has a massive negative impact on the overall sum throughput of eMBB users. This is the main reason why we opt for considering a minimum data rate for eMBB users instead of forcing a fairness level. This idea can protect eMBB UEs at the cell edge from starvation and provide them with acceptable data rates. Moreover, this method would also elevate the sum throughput by not limiting the achievable data rates of eMBB UEs with good channel conditions.

Constraint (7a) can be transformed into a deterministic form using CDF which helps in avoiding the complexity that comes along with any stochastic variable. This method can be quite inefficient as the deterministic form can sometimes be very complex depending on the random distribution from which the random variable is derived and the CDF of this distribution. In our case, a Pareto distribution is used to produce the uRLLC load. This would enable us to work with a relatively simple CDF outcome that can be easily relaxed in our optimization process. The idea is that if X is a Pareto random variable, we can calculate the probability that X is greater than a value x. The CDF of the Pareto distribution is given as follows:

$$F_X(x) = \begin{cases} 1 - \left(\frac{x_m}{x}\right)^{\alpha} & x \geq x_m \\ 0 & x < x_m \end{cases} \tag{8}$$

where x_m is the minimum positive value of x and represents the scale parameter of the Pareto distribution. α is a positive value that represents the shape parameter of the Pareto distribution. We can apply (8) on the constraint (7a) as shown below: let us assume that the term u represents the outage probability of the uRLLC users.

$$u = P\left[\sum_{j=1}^{m} B_{iu}(j)\log_2(1 + SNR_u(j)) < L_u\right] \leq \gamma \tag{8a}$$

Then, we apply (8) to u as shown below.

$$P[u < L_u] \leq \gamma \Leftrightarrow 1 - F_X(u) \leq \gamma \tag{8b}$$

$$\Leftrightarrow F_X(u) \geq (1-\gamma) \tag{8c}$$

$$\Leftrightarrow u \geq F_X^{-1}(1-\gamma) \tag{8d}$$

$$\Leftrightarrow u^{\alpha} \geq \frac{x_m^{\alpha}}{\gamma} \tag{8e}$$

Here, $F_X^{-1}(1-\gamma)$ is the inverse CDF of uRLLC load which is evaluated using the reliability level defined earlier that simply ensures the delivery of the uRLLC load with its latency budget regardless of the payload size. Equation (8a) shows how the uRLLC random payload size is transformed into a deterministic form based on a predefined reliability level γ. As a result, the constraint in (7a) can be redefined as follows:

$$\left(\sum_{j=1}^{m} B_{iu}(j)\log_2(1 + SNR_u(j))\right)^{\alpha} \geq \left(\frac{x_m}{\gamma}\right)^{\alpha} \tag{9}$$

where x_m and α are the scale and the shape of the Pareto distribution respectively. The formulation is now following a convex form and thus a global maximum can be achieved.

3.2 Sub-optimal resource allocation of eMBB/uRLLC traffic

The optimal allocation has been addressed intensively in the literature where most of the time the solution is with high complexity. The uRLLC traffic as indicated earlier has a stochastic nature and strict requirements and needs to be served instantaneously. This is one of the reasons why low-complexity solutions are considered more practical even with their lower efficiency when compared to the optimal approaches. At the BS, the resource allocation decision must be taken immediately, and the need for complex calculations makes it difficult to cope with the traffic density and satisfy the diverse requirements of different services. In this part, we address eMBB-aware scheduling algorithms for uRLLC with each having a different objective. All these algorithms are based on the resource puncturing approach in which the uRLLC is instantly served upon arrival. The main idea is to test different resource puncturing algorithms that would provide the best performance possible for both eMBB and uRLLC.

The puncturing process is vital in determining the level of impact on every user. Moreover, the decision of RB selection has a crucial role in elevating the efficiency of the puncturing algorithm. Different parameters are considered in making the puncturing decision upon uRLLC traffic arrival. The channel conditions of both eMBB and uRLLC users represent the most important factor in the decision as it affects the users' data rates directly since each user might experience different channel quality at each RB. These channel conditions are estimated by the BS through channel state information feedback, that is sent regularly by the UEs to the BS. It is important to consider the state of the user at these RBs before puncturing in order to preserve fairness among eMBB users, provide better reliability for uRLLC, and to maximize the data rate of each eMBB user.

3.2.1 Best resource block for uRLLC (Algorithm 1)

The objective of the first algorithm is to provide the best possible reliability level for uRLLC traffic considering the channel condition of the selected uRLLC UE. This is done by allocating RBs with the best channel condition of the selected uRLLC UE. This algorithm not only provides uRLLC UE with better reliability levels, but also prevents the puncturing of extra resources in order to satisfy the latency requirements of the uRLLC traffic. In fact, better channel conditions mean a higher Modulation and Coding scheme value can be assigned to the uRLLC UE and thus more data can be transmitted using fewer resources. Slot boundary is taken into account and the algorithm is updated once the RB is entirely consumed, moving to another RB where the uRLLC UE channel condition is the best compared to the other available RBs.

3.2.2 Protecting eMBB UE at the cell edge (Algorithm 2)

The second algorithm aims to protect the eMBB UE at the edge level in order to prevent their starvation. Users at the cell edge most likely suffer from bad channel conditions and cannot generally tolerate the effect of puncturing their resources. The CQI of each user is an important indicator that would help the BS to distinguish and apply protection policies that would lower the impact on these users. Protecting those users can be achieved by allowing the resource puncturing of eMBB UE with the best channel conditions as these users can be less affected by low uRLLC traffic density and their QoS level can be maintained even with the presence of uRLLC.

3.2.3 Maximization of eMBB sum-throughput (Algorithm 3)

The third algorithm aims to maximize the sum throughput of eMBB UEs while satisfying the requirements of uRLLC UEs. This can be achieved by targeting the resources of eMBB UE with lower channel conditions in order to protect the eMBB users with higher contributions to the overall sum throughput. It can be noticed that the previous two algorithms might target the same eMBB UE in the case of having a large uRLLC payload size or having multiple uRLLC transmissions at the same time slot.

For Algorithms 1 and 2, the data rate of this eMBB UE at the punctured RB is updated according to Eq. (2). The channel quality is based on the *SNR* level of the UE over its assigned subcarrier. In this paper, we used the approach proposed in [37, 38] to calculate the CQI value of UE as a function of the *SNR* values of the selected user over all its assigned sub-carriers. It is important to mention that the CQI reporting by UE is assumed to occur every 1 ms (1 Time Slot) which is vital for the algorithms to operate efficiently.

The time complexity is the same for all three algorithms which is in the order of $O(n^2)$ where n is the number of uRLLC UEs in the case of Algorithm 1 or the number of eMBB UEs in the case of Algorithms 2 and 3. Here we are performing a linear search for a maximum or minimum CQI value with $2(n-1)$ comparisons at each time slot, assuming a worst-case scenario with continuous incoming uRLLC traffic.

Given a number of resource blocks RBs depending on the used numerology and a time slot size of 7 minislots each consisting of 2 symbols, a summary of the above reference algorithms is shown below (Table 1).

Table 1 Algorithms 1, 2 and 3 parameters

Parameter	Meaning
E	eMBB UE with allocated RBs at TTI t
U	uRLLC UE demanding immediate service
CQI_E	Array of eMBB UE CQI values in each RB at TTI t
CQI_U	Array of uRLLC UE CQI values in each RB at TTI t
D_{size}	Payload size of uRLLC UEs
R_{uk}	Data rate of uRLLC user u at RB k
N_{TTI}	Number of TTIs
O_{RB}	Algorithm Output = Selected RB for puncturing
N_{RB}	Number of available RBs
N_{mini}	Number of mini-slots

```
Algorithm 1, 2 and 3
Inputs: $E$, $U$, $CQI_E$, $CQI_U$, $D_{size}$, $N_{RB}$
Outputs: $O_{RB}$,
for $TTI = 1$ to $N_{TTI}$ do
    Schedule eMBB UE using PFTF, BCQI, or M-LWDF for all RBs
    $N_{mini} = 0$
    while (any ($D_{size}$ > 0) && $N_{RB}$ > 0) do
        Select a uRLLC UE to serve on a first come first served basis
        switch Algorithm do
            case Algorithm 1
                Select $RB_k = argmax\ (CQI_U)$
            case Algorithm 2
                Select $RB_k = argmax\ (CQI_E)$
            case Algorithm 3
                Select $RB_k = argmin\ (CQI_E)$
        $N_{mini} = N_{mini} + 1$
        if $N_{mini} == 7$ then
            $N_{RB} = N_{RB}$ - 1
            $N_{mini} = 0$
        end if
        Allow Puncturing
        Update $D_{size}$ using Eq. (2) $\rightarrow$ $D_{size}$= $D_{size}$ - $R_{uk}$
    end while
end for
```

3.2.4 Knapsack-inspired uRLLC fair punctured scheduling (Algorithm 4)

The fourth algorithm is a knapsack-inspired scheduling algorithm which aims to maximize the sum throughput of the eMBB UEs while satisfying the requirements of the uRLLC traffic, and preserving a fairness level among the eMBB UEs in terms of the amount of punctured resources from each of these users. This algorithm includes a number of objects representing the RBs in which each object has a profit and a weight associated with it. In each RB, the weight represents the data rate of the eMBB UE occupying it, and the profit is the channel condition of the uRLLC UE at this RB. Better channel condition means that more data are being sent using this RB and thus more profit is gained. The channel condition is measured using the SNR of the selected uRLLC UE. The weight reflects the amount of impact on the sum throughput of eMBB UE upon puncturing this RB. The knapsack is the constraint that needs to be considered when solving the problem and it represents the payload size of the uRLLC UE that needs to be transmitted within the current time slot. The objective is to fill the knapsack in a way that maximizes the profit while considering the constraint. The solution is given in a form of a set where each element shows if the RB has been selected for puncturing or not. The element value is between 0 and 1 which means that the RB can be partially punctured depending on the number of mini-slots given to uRLLC UE. In order to select the most suitable RB that leads to the best profit while considering its weight, we need to take the profit-by-weight ratio. To do that, channel conditions of uRLLC UE and the achieved data rates of eMBB are rescaled in the range of 1–100 (this is to avoid any issues as the two parameters have different ranges). The RB with the highest profit by weight ratio is selected where a fraction of this RB is punctured and added to the solution set. The fraction of the RB represents the TTI duration of the uRLLC or the size of the mini slot (2, 4, 7 OFDM symbols). After each selection of RBs, the payload size (knapsack) is updated based on the amount of data that we were able to transmit using this RB. This depends on the channel condition of the selected uRLLC UE at this RB. It can be noticed that one resource block can be entirely targeted throughout the whole process and starvation of certain eMBB UE is expected once their RB has a higher profit/weight ratio. To prevent this and to provide a sense of fairness, we included a second constraint to the problem in which the resource block cannot be punctured two times in a row and the algorithm will move to another RB representing the second highest profit/weight ratio until the entire uRLLC payload is transmitted.

At time slot t, each algorithm analyzes all RBs in order to find the most suitable one according to its criteria. All uRLLC UEs are considered to be using the same type of application where the latency requirement is 1 ms and all have the same priority.

The time complexity of the proposed knapsack problem is in the order $O(n \log n)$ which is acceptable in practical implementation. The time complexity was analyzed for each TTI with respect to the resource block assignment process in addition to the UE selection which is based on a standard sorting algorithm. The knapsack-inspired algorithm is summarized as Algorithm 4, shown next, with its parameters listed in Table 2.

Algorithm 4 Knapsack inspired resource allocation scheme for eMBB and uRLLC traffic

```
Inputs: $E$, $U$, $SNR_U$, $SNR_E$, $D_{size}$
Outputs: $O_{RB}$,
$idx$=1;
for $TTI = 1$ to $N_{TTI}$ do
    while (any ($D_{size}$ > 0) && any ($P$ < $N_{mini}$)) do
        Select uRLLC UE ($U$)
        Calculate $R_{ek}$ in each RB using Eq. (2) then store in $R_{inst}$
        Calculate the ratio of $SNR_U$ & $R$ in each RB then store in $Ra$
        Sort $Ra$ in descending order
        if $P < N_{mini}$ then
            Puncture RB at $Ra(idx)$
            $P$=$P$+2
            $idx$=$idx$+1;
        end if
        Update $D_{size}$ using Eq. (2) $\rightarrow$ $D_{size}$= $D_{size}$ - $R_{uk}$
        if $idx > N_{RB}$ then
            $idx$=1;
        end if
    end while
end for
```

4 Simulation results

In this section, all the results of our simulations are presented in addition to a detailed analysis of the performance of our proposed algorithm. All simulations were done using MATLAB® with Intel(R) Core(TM) i7-4700MQ CPU @ 2.40 GHz and 16 GB RAM. To evaluate the performance, we induced a set of simulations that involved a varying number of eMBB UEs randomly distributed around the base station and scheduled according to several schedulers defined in Sect. 2. It is assumed that all the resources have been previously allocated to eMBB UEs. All simulations included a minimum uRLLC load size of 1 Mbps assuming a minimum uRLLC packet size of 32 bytes [31]. eMBB traffic represents a real-time video streaming application with an average delay threshold of 100 ms and a packet loss ratio threshold of 10%. Resource puncturing is restricted to slot boundary and cannot exceed the following slot. The algorithms are applied in each TTI and evaluated in terms of Throughput, Fairness, and Spectral Efficiency. The evaluation metrics can be calculated using the below equations. Note that the data rate can be calculated using Eq. (2). The first evaluation metric is the fairness of the scheduler which can be measured using the well-known, Jain's fairness index [39] that can be used to determine if each user is receiving an equal share of resources compared to others.

$$\text{Fairness Index} = \frac{[r_i]^2}{N \sum_{i=1}^{N} r_i^2} \tag{10}$$

where N is the number of eMBB users and r_i is the data rate of eMBB user i. In our simulations, we are adopting a full buffer model in which the eMBB users will always have data to transmit resulting in full resource usage. Nevertheless, the amount of resources allocated to each user will differ according to each scheduler, thus resulting in different amounts of bits transmitted, because of their

Table 2 Algorithm 4 parameters

Parameter	Meaning
E	eMBB UE with allocated RBs at TTI t
U	uRLLC UE demanding immediate service
SNR_U	Array of SNR values of uRLLC UE over all RBs
SNR_E	Array of SNR values of eMBB UE over all RBs
D_{size}	Payload size of uRLLC UEs
R_{inst}	Array of eMBB UE instantaneous data rates at TTI t
R_{uk}	Data rate of uRLLC user u at RB k
R_{ek}	Data rate of eMBB user e at RB k
Ra	Array of eMBB data rates and uRLLC SNR ratios
P	Record of the amount of punctured resources from each eMBB UE
idx	Index of the last user with punctured resources
N_{TTI}	Number of TTIs
O_{RB}	Algorithm Output = Selected RB for puncturing
N_{RB}	Number of available RBs
N_{mini}	Number of mini-slots

different objectives and the channel condition each user is experiencing. This is why the second metric is important to measure how efficiently the bandwidth is used. Spectral efficiency [40] is used to measure the data rate that can be transmitted in a specific bandwidth through a cellular network and can be calculated using the following equation:

$$\text{Spectral Efficiency} = \frac{\text{Sum throughput}}{\text{Total BW}} \tag{11}$$

The simulation parameters are listed in Table 3. The first evaluation is done on Algorithm 1, in which the RB where uRLLC UE experiences the best channel condition is selected for puncturing with the main goal of providing a high-reliability level for uRLLC and avoiding the need for re-transmissions in case of lost packets. This algorithm has a big advantage in that the better uRLLC channel condition would result in more data being sent at this resource block and thus, fewer resources are needed by the uRLLC traffic to satisfy its requirements which lowers the negative impact on the performance of eMBB. Nevertheless, this approach does not consider fairness among eMBB UEs and can be considered as an eMBB unaware approach because it is possible that certain RBs will be entirely consumed by the uRLLC UE and those eMBB UEs occupying these RBs will be impacted more than others. When comparing the proposed algorithms, we investigate the level of degradation in each of the evaluation metrics when allowing uRLLC traffic with different densities to puncture the resources of eMBB UEs. The simulation of the two scenarios (with and without uRLLC traffic) has been performed at the same time which is important to make sure that the channel conditions are the same in both scenarios. The number of uRLLC UEs increases after every simulation run.

Figure 2, shows the sum throughput of eMBB UEs when impacted by uRLLC traffic. We can notice the huge drop in the performance of Best CQI and M-LWDF as these algorithms consider the QoS of the served users which is (in the simulation settings) related to the channel quality of the user over all of the user's assigned resource blocks. QoS requirements are often linked to the channel condition of the user and the distance from the BS, and we can observe that those users were targeted by the puncturing algorithm leading to a larger degradation in the overall sum throughput of eMBB. It can be noticed that the fair resource allocation by PFTF lowered the impact on eMBB users at some level, given that most of those users had a similar amount of resources allocated to them and the algorithm was not biased by a single metric during the resource assignment. Figure 3 shows the same effect of uRLLC resource puncturing in terms of spectral efficiency which is directly related to the amount of achieved data rate and we can notice that all three algorithms recorded about 40% decline with the increase of uRLLC traffic density (i.e., increase in the number of users and thus increase in the total traffic size).

Algorithm 2 targets eMBB users with the highest achieved data rate across all RBs in time slot t. The main goal is to protect eMBB UEs at the cell edge who most likely suffer from bad channel conditions and thus cannot tolerate the puncturing of their resources. Figure 4 shows the performance of Algorithm 2 on the sum throughput of eMBB. The degradation is severe, and the performance is much worse compared to Algorithm 1 discussed earlier, as it tends to target eMBB UEs with high data rates (i.e., best channel conditions). This directly affects the sum throughput as these users have more contribution to the sum throughput than others. The main advantage is the prevention of starvation

Table 3 Simulation parameters

Parameter	Value
BS max power/UE	21 dBm
Cell radius	1 km
Total bandwidth	10 MHz
MIMO	2 × 2
Propagation model	$-128.1 + 37.6 * log_{10}(d) \rightarrow$ d: UE Distance from BS in km
Channel	Rayleigh distributed fading
Number of eMBB users	50
UE distribution	Randomly distributed
eMBB traffic model	Full buffer
CQI reporting	Every 1 ms
UE noise figure	7 dB
Number of slots	1000
Sub-carrier spacing	15 kHz
Time slot size/duration	14 symbols/1 ms
Slot format	0 [13]

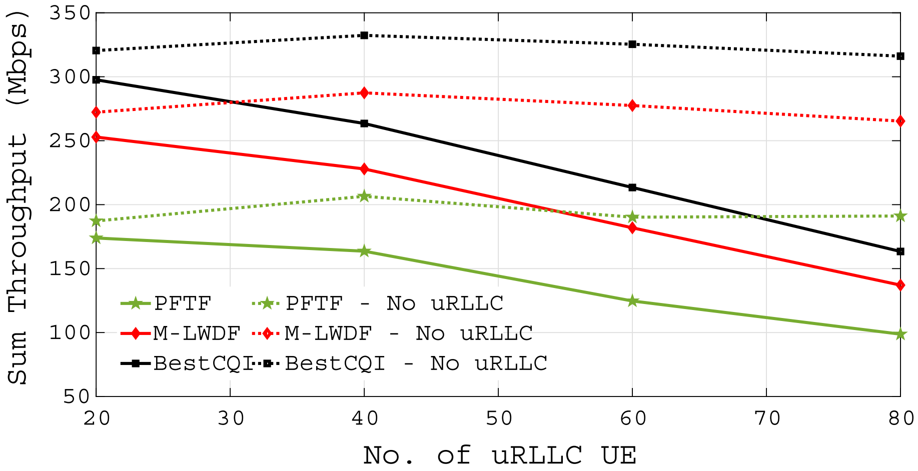


Fig. 2 eMBB sum throughput versus No. of uRLLC UE (Algorithm 1)

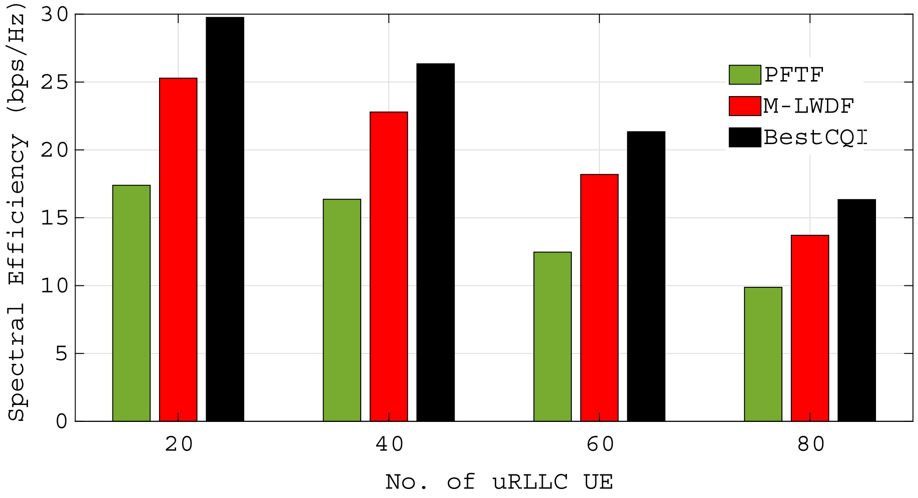


Fig. 3 eMBB spectral Eff. versus No. of uRLLC UE (Algorithm 1)

and although the data rate of eMBB declined, those users still have connectivity and can transmit data via their remaining resources where they experience good channel condition. In other words, those users are more likely to experience good channel conditions on their remaining RBs and thus can transmit a good amount of data because of this advantage.

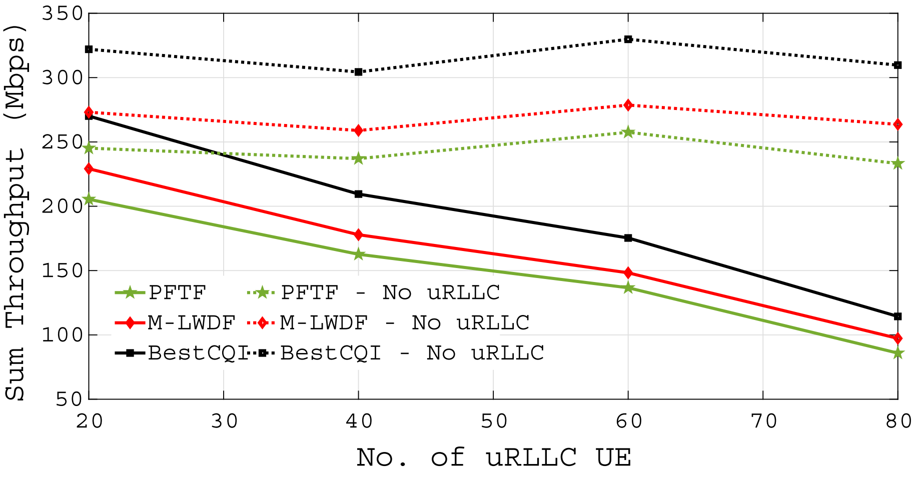


Fig. 4 eMBB sum throughput versus No. of uRLLC UE (Algorithm 2)

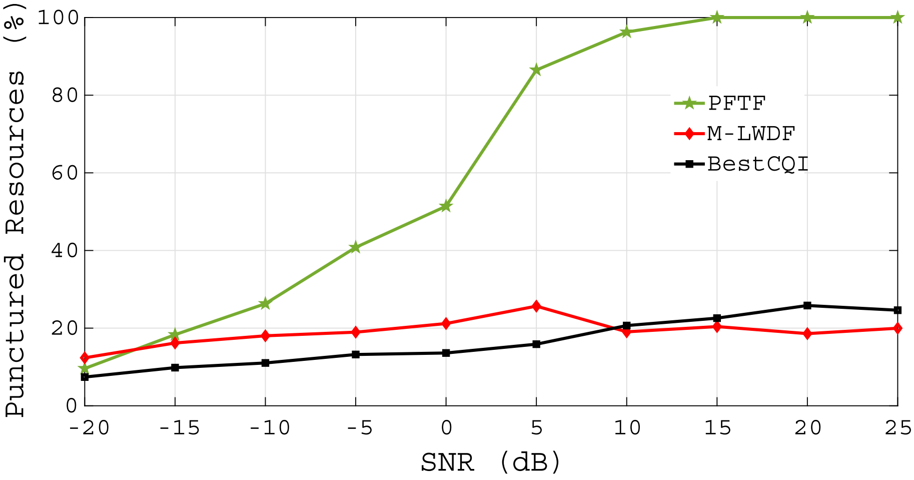


Fig. 5 eMBB SNRs versus % of punctured resources (Algorithm 2)

Figure 5, shows the effectiveness of this algorithm in protecting eMBB UEs with bad channel conditions. As the figure shows, the percentage of punctured resources from those users is much less than those with better channel conditions.

Algorithm 3 aims to maximize the data rate of eMBB UEs without considering the reliability level given to uRLLC and the fairness among eMBB UEs. The behavior of this algorithm tends to target eMBB UEs with low channel conditions (i.e., have a low contribution to the overall sum throughput) aiming to protect eMBB UEs with high data rates. This would not only raise the sum throughput but also the spectral efficiency. Nevertheless, the algorithm does not take into account the possibility that these uRLLC UEs might be experiencing bad channel conditions on the allocated RB, and thus the need for more resources increases in order to satisfy the latency requirements of uRLLC traffic.

Figure 6 shows the sum throughput of eMBB UEs and we can notice that the degradation is less when compared to the previous two algorithms. This is expected given the main objectives of the algorithm. Nevertheless, the efficiency of this algorithm can be considered acceptable when the evaluation is on the system level and the aim is to improve the overall performance without considering the state of the users and how they are affected individually.

Figure 7, shows the percentage of punctured resources of 10 eMBB UEs where we can notice that the behavior of each of the heuristic scheduling algorithms is reflected in the result. The BestCQI and M-LWDF algorithms have almost a uniform percentage of punctured resources among all 10 users. This is explained by the fact that these algorithms provide more resources for users with good channel conditions and thus those users are not being targeted directly by the puncturing algorithm.

For our proposed algorithm, the main objective of knapsack inspired algorithm is to find the best resource block for puncturing by which we maintain an acceptable level of fairness in addition to achieving the best possible sum throughput of eMBB UEs. This algorithm considers the channel conditions of both eMBB and uRLLC UEs which gives it the privilege over the

Fig. 6 eMBB sum throughput versus No. of uRLLC UE (Algorithm 3)

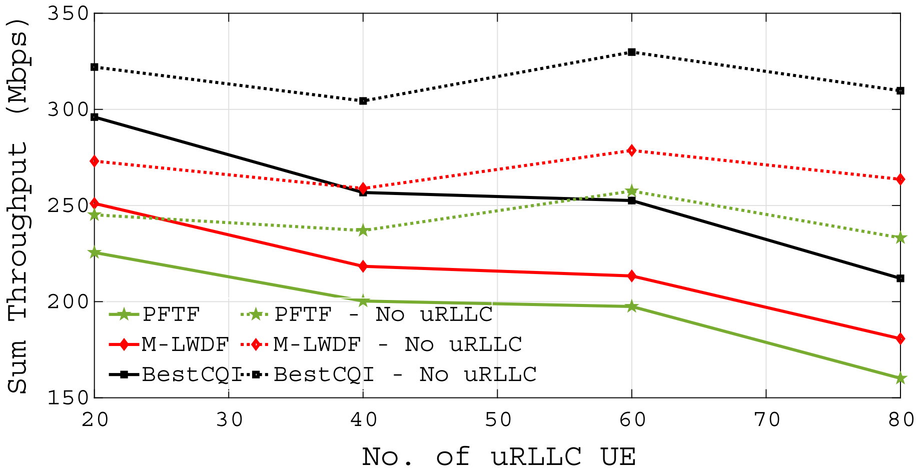


Fig. 7 Percentage of punctured resources among eMBB UEs (Algorithm 3)

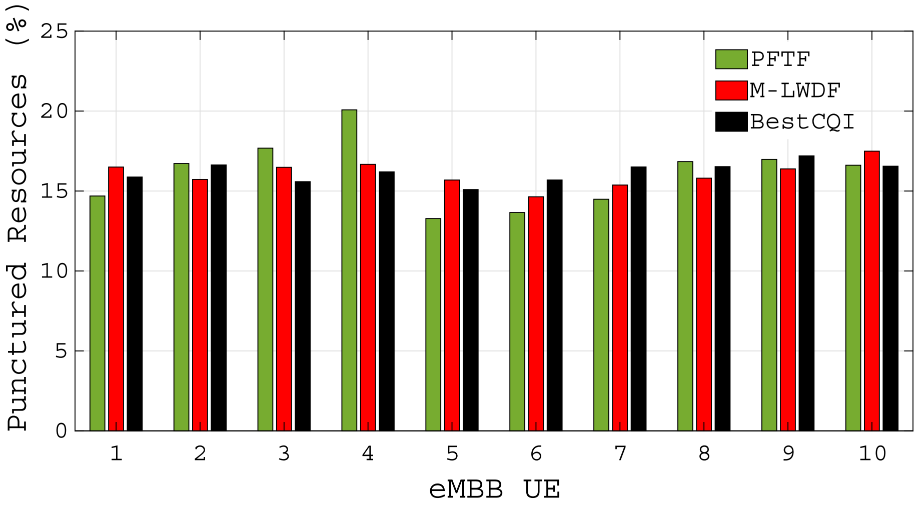
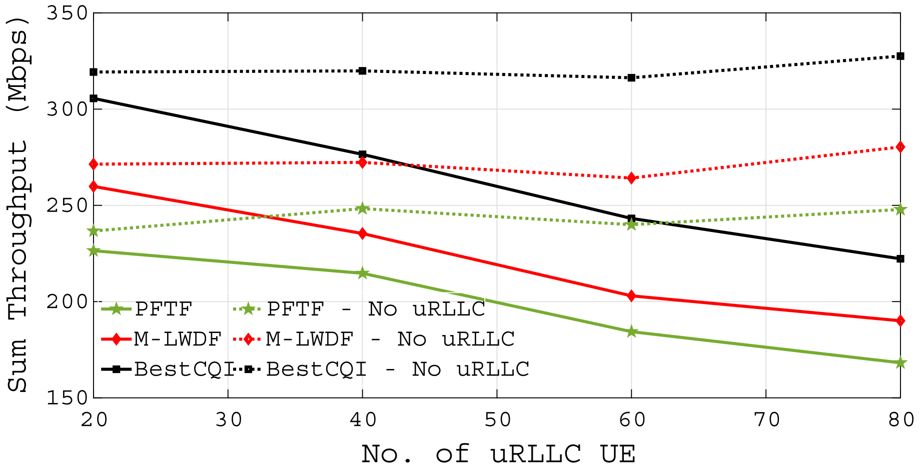
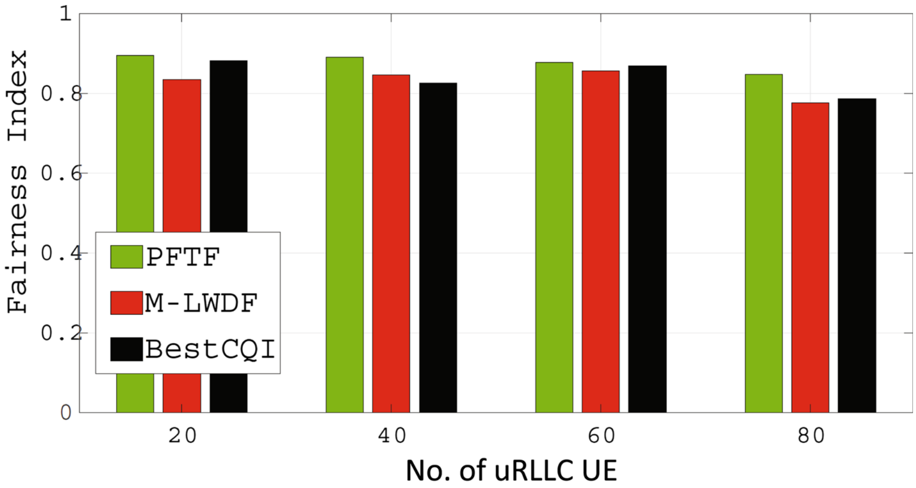


Fig. 8 eMBB sum throughput versus No. of uRLLC UE (Algorithm 4)



previously discussed algorithms. It simply tackles the limitations found in the discussion of our results above, trying to balance the trade-offs between fairness level and sum throughput. Figure 8 shows the performance of this algorithm as it achieves the lowest degradation level in eMBB sum throughput when punctured by an increasing number of uRLLC UEs, compared to the previous algorithms. This is done by utilizing the knowledge of channel conditions of eMBB and uRLLC UEs in a way that leads to the selection of the most suitable RB for puncturing that

Fig. 9 Fairness index versus No. of uRLLC UE. (Algorithm 4)



would benefit both eMBB (does not degrade the sum throughput) and uRLLC (provides an acceptable level of reliability).

Figure 9 shows the puncturing of eMBB users by different numbers of uRLLC users. This graph is a strong indicator of the effectiveness of our proposed algorithm (Algorithm 4), in terms of forcing a fairness level among affected eMBB UEs. This can be noticed in the sudden change in the behavior of M-LWDF and BestCQI. The figure shows that their fairness level is almost similar to that of the PFTF algorithm. The fairness index is measured based on the individual data rate compared to others. This can be clarified by pointing out that the algorithm considers the achieved data rate of eMBB users in every RB, and after each puncturing process which takes 2 OFDM symbols, this RB is also a potential target for puncturing in the next uRLLC allocation. Nevertheless, the algorithm forces the transition to another RB with the highest profit/weight ratio to enforce a level of fairness in terms of the amount of punctured resources. Also, the data rate of the last punctured eMBB user is updated which lowers the profit in this RB and decreases the possibility of selecting it in future uRLLC allocations. This led to a fair percentage of punctured resources from each eMBB user depending on the user's situation at the time of puncturing.

Figure 10 Shows the percentage of punctured resources and reflects the level of fairness even more clearly as it illustrates an almost fixed percentage of punctured resources among 10 eMBB UEs. The simulation included 10 eMBB UEs and 20 uRLLC UEs in order to reflect a 6 G-like scenario where the BS might encounter a large amount of uRLLLC traffic, even double the size of eMBB traffic. The enforced fairness level provides a stable service to these users which is an important factor in evaluating the performance of every algorithm.

In the rest of this section, we consider an optimal allocation of resources conditioned by obtaining the desired level of fairness among eMBB UEs. This was done by adding Jain's fairness index as a constraint in the optimization problem formulation in order to push the optimizer to provide fair allocation by a predefined level. The simulation included an equal number of eMBB and uRLLC UEs (20 users) which is a more realistic scenario for a 6 G environment and the performance was measured on a slot basis. Figure 11 shows the impact of different fairness index values forced on the optimizer and how it affects the sum throughput of the eMBB UEs while being punctured by uRLLC traffic. It can be noticed that the higher the fairness level required, the larger the impact is on the sum throughput of eMBB UEs. This is because the optimization algorithm can no longer select eMBB UEs with bad channel conditions (i.e., less contribution to the overall sum throughput) for puncturing, and it is forced on fair treatment of connected UEs according to the set fairness index value. The fairness index compares the achieved throughput of each eMBB UE with the other users in order to verify the fairness of the system by observing the difference between each user's data rate.

The next part of the simulation results is used to evaluate the effect of specified uRLLC outage probability on the eMBB sum throughput. Figure 12 shows how the sum rate declines with the decrease in the outage probability. A low outage probability provides a higher reliability level for uRLLC UE in a way that enforces the optimization algorithm to assign better resources for those uRLLC users. It can be seen that the increase in the tolerated outage probability improves the performance in terms of eMBB sum throughput. This can be justified by the fact that the BS is no longer forced to allocate the best RB for the uRLLC in order to maintain the required reliability level. This might enhance the performance of eMBB users in the case where the RB of the uRLLC (with better channel conditions) is the same RB used by an eMBB UE with a large contribution to the sum throughput.

Fig. 10 Punctured resources % among eMBB UE (Algorithm 4)

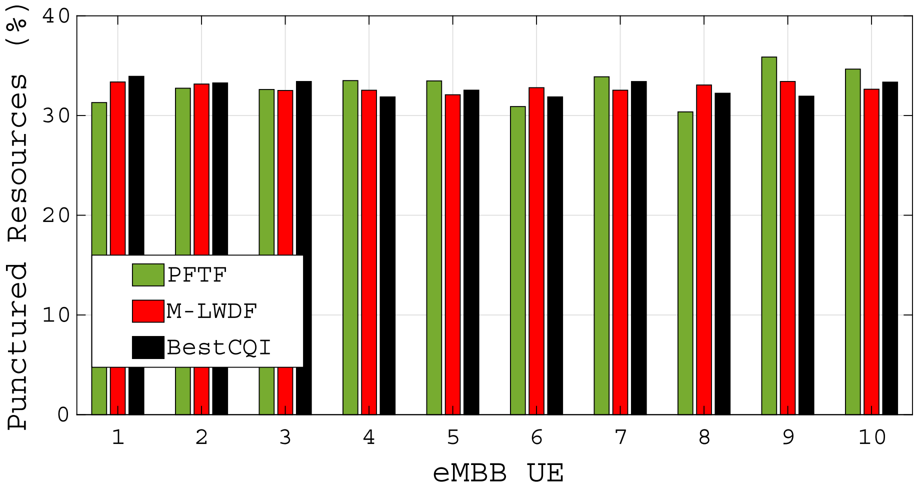


Fig. 11 eMBB throughput versus fairness levels

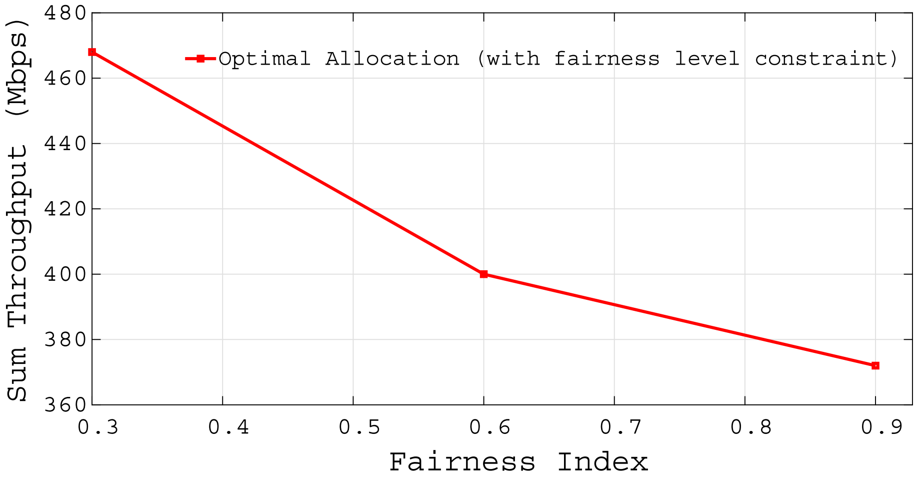


Fig. 12 eMBB throughput versus outage probability (with 0.5 fairness level condition)

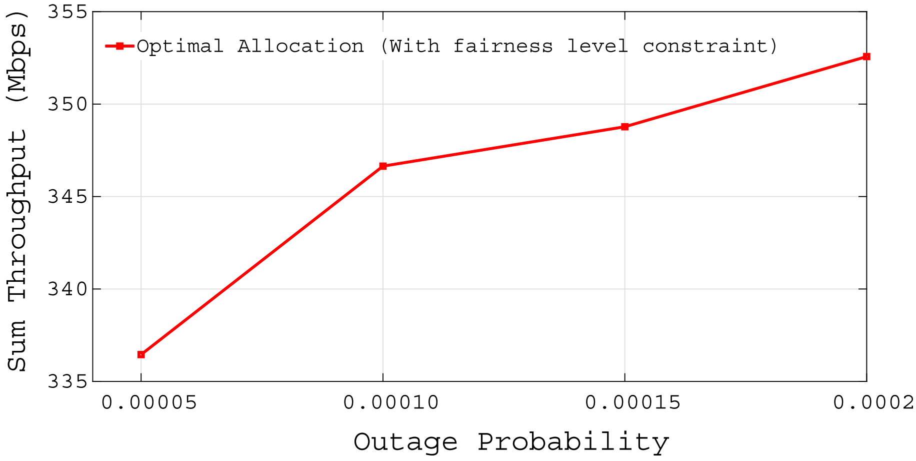


Figure 13 shows how the proposed optimal resource allocation scheme is affected by different uRLLC traffic densities. The degradation of the eMBB throughput is sharp but the overall achieved sum throughput is much larger than the sub-optimal approach.

Figures 14 and 15 show the results of the optimal allocation while adding the condition of the eMBB user's minimum data rate. Figure 14 shows the sum throughput of eMBB UEs when different outage probability values are tolerated by uRLLC. We can notice that the result is better than the result in Fig. 12, where a fairness level among all

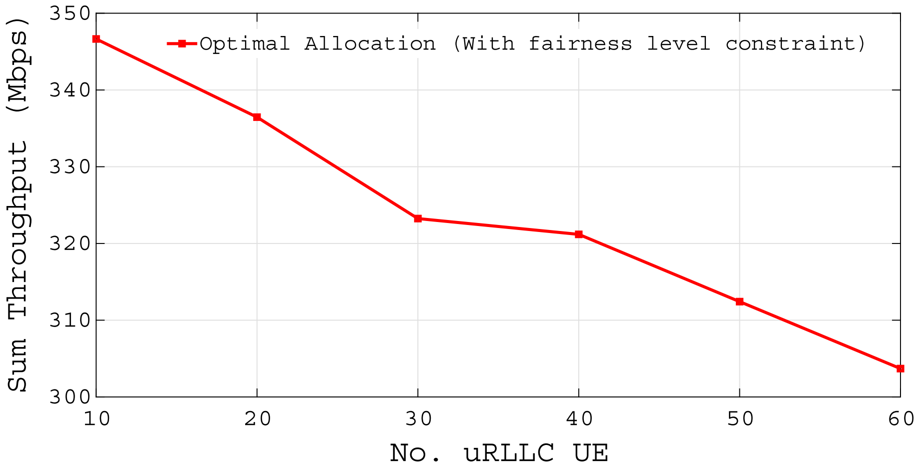


Fig. 13 eMBB throughput versus No. uRLLC UE (with 0.5 fairness level condition)

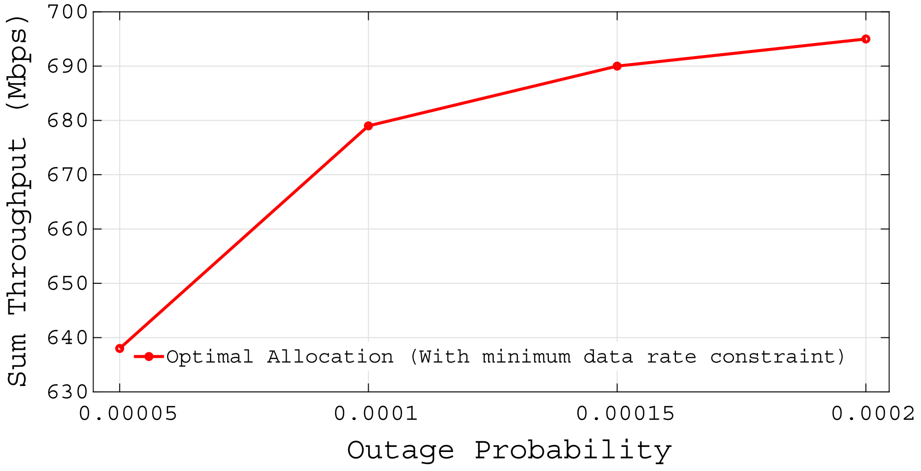


Fig. 14 Outage probability versus eMBB sum throughput (with 11 Mbps/eMBB UE minimum data rate condition)

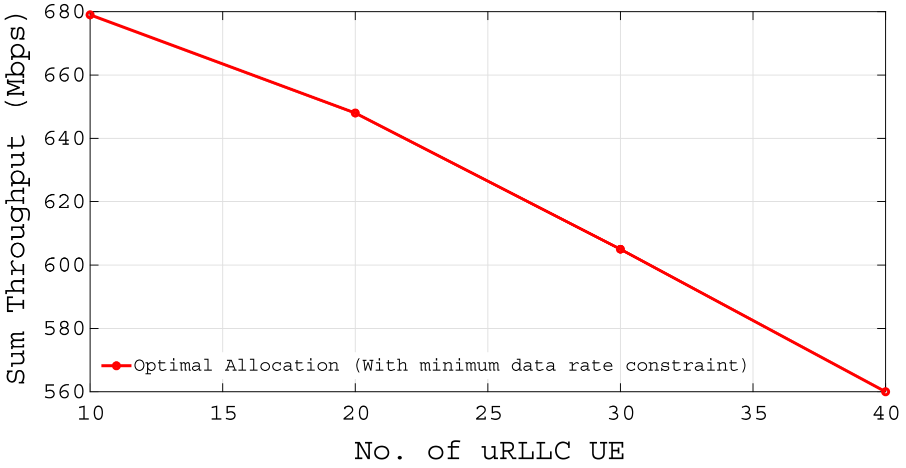


Fig. 15 eMBB sum throughput versus No. of uRLLC UE (with 11 Mbps/eMBB UE minimum data rate condition)

users was required. This is due to the fact that the eMBB UEs with bad channel condition might have a huge gap in terms of data rate when compared to eMBB users with good channel condition. Adding a constraint that aims to narrow this gap will result in allocating resources to those UEs with bad channel conditions in order to try to achieve the ultimate fairness of equal rate, resulting in a higher drop of eMBB sum-throughput.

Figure 15 shows the impact of different numbers of uRLLC UEs on the sum throughput of eMBB UEs. The result includes the condition of maintaining a minimum data rate of 11 Mbps for each eMBB UE. This shows the

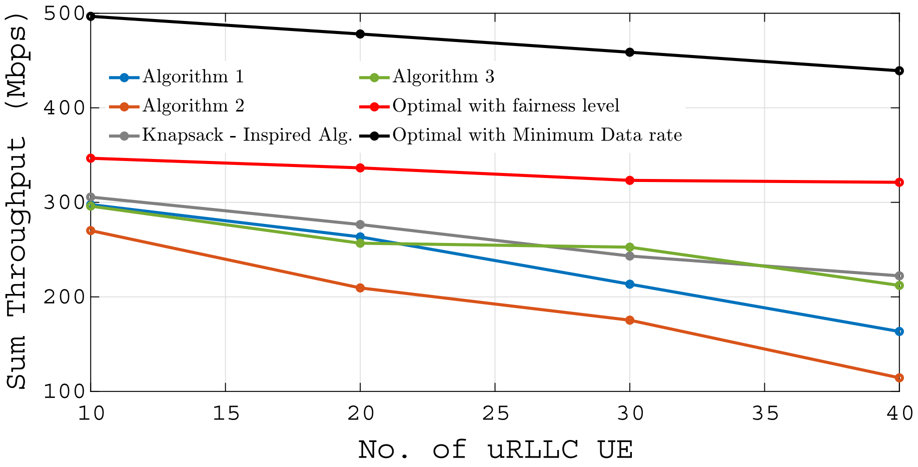


Fig. 16 Sub-optimal versus optimal resource allocation

advantage of adding this constraint instead of the fairness index as shown in Fig. 13. The minimum data rate condition is more realistic and feasible in most scenarios that could include low, medium, or high uRLLC user densities. The degradation is acceptable given that the maximum achieved sum throughput is higher than the one in Fig. 13.

Finally, we compare the four sub-optimal allocation algorithms to the optimal allocation algorithms in order to observe the gap between these two types of allocations. The sub-optimal allocation included the reference algorithms, Algorithm 1, which aims to provide the best resource block for uRLLC users, Algorithm 2, which aims to protect eMBB users at the cell edge, Algorithm 3, which aims to maximize eMBB sum throughput without considering any other metrics and lastly, our proposed solution, Algorithm 4 which is a knapsack inspired algorithm that aims to maximize eMBB sum throughput while maintaining a level of fairness. The Optimal allocation included two parts where the first aims to maintain a predetermined fairness level and the second one aims to guarantee a minimum data rate for eMBB users.

Figure 16 shows a performance comparison among all proposed algorithms. It is clear that the optimal allocation methods reach a higher level of performance compared to the sub-optimal ones but the fact that these solutions require intensive processing capability by the BS, due to their high computational complexity, reduces their practicality compared to sub-optimal solutions. It can be noticed that the knapsack-inspired algorithm provides acceptable performance with a key feature of low complexity which demonstrates its effectiveness and feasibility for real-life implementation. It outperforms all the reference sub-optimal allocation algorithms in terms of eMBB sum throughput. Algorithm 1 provided a better performance with low user density compared to Algorithm 3. Algorithm 2 achieved the worse performance given that it targets eMBB users with the highest contribution to the sum throughput. The optimal allocation with guaranteed minimum throughput provided the best results compared to all others.

5 Conclusion

In this paper, we addressed the eMBB and uRLLC resource allocation problem in a 6 G-like Scenario. Two approaches have been proposed. The first one includes an optimal resource allocation between eMBB and uRLLC services in addition to addressing the optimization under uncertainty. The formulation is based on transforming the stochastic uRLLC traffic into its deterministic form aiming to maximize the eMBB data rate while not exceeding a pre-determined outage probability threshold. The approach also aims to satisfy the desired fairness level among eMBB UEs in terms of the percentage of punctured resources which is vital in protecting users with bad channel conditions. The second approach includes a sub-optimal solution to the problem that features low complexity and acceptable performance in terms of achieved eMBB sum throughput and fairness level. The approach consists of a puncturing method that aims to use the knowledge of the users' channel conditions in order to make an optimal selection of RBs prior puncturing phase. The problem formulation is a knapsack-inspired formulation in which the ratio of eMBB achieved data rates at each RB and the CQI of the uRLLC UE at each RB is used as a decision parameter to maximize the eMBB sum throughput while satisfying the requirement of uRLLC and providing the best possible reliability level at each time slot. A set of simulations have been conducted with uRLLC traffic of different intensities arriving stochastically. The other sets of simulations included uRLLC traffics with a minimum data rate of 1 Mbps. The latency and reliability requirements of uRLLC have been considered and strongly enforced when served by the BS.

We showed that the proposed algorithm can minimize the impact of uRLLC traffic on the schedulers' performance in terms of Sum throughput, Fairness, and Spectral efficiency.

Acknowledgements The authors would like to thank the Editor and the anonymous reviewers for their valuable comments. This publication was jointly supported by Qatar University and IS-Wireless-IRCC Grant No. IRCC-2021-003. Open Access publication fees were paid by Qatar National Library (QNL). The findings achieved herein are solely the responsibility of the authors.

Funding Open Access funding provided by the Qatar National Library.

Data availability Data sharing is not applicable to this article as no datasets were generated or analysed during the current study.

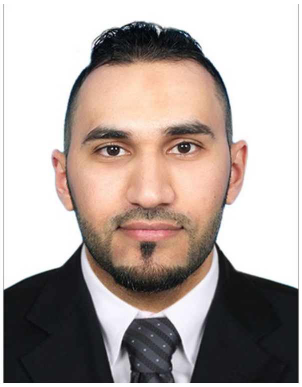
References

1. Ericsson. (2022). Ericsson mobility report.
2. Morgado, A., Huq, K. M., Mumtaz, S., & Rodriguez, J. (2017). A survey of 5g technologies: Regulatory, standardization and industrial perspectives. *Digital Communications and Networks, 4*(2), 87–97.
3. ITU-R: Imt vision-framework and overall objectives of the future development of imt for 2020 and beyond. Recommendation m.2083-0 (September 2015). International Telecommunication Union.
4. Zaidi, A., Athley, F., Medbo, J., Gustavsson, U., Durisi, G., & Chen, X. (2018). Introduction: 5g radio access, 1–19, Chapter 1 in the Book: *5G Physical Layer, Principles, Models and Technology Components*, Academic Press, ISBN: 978-0-12-814578-4.
5. Osseiran, A., Monserrat, J. F., Marsch, P., Queseth, O., Tullberg, H., Fallgren, M., Kusume, K., Höglund, A., Droste, H., Silva, I., Rost, P., Boldi, M., Sachs, J., Popovski, P., Gozalvez-Serrano, D., Fertl, P., Li, Z., Sanchez Moya, F., Fodor, G., & Lianghai, J. (2016). 5G mobile and wireless communications technology.
6. 3GPP: Study on physical layer enhancements for nr ultra-reliable and low latency case (urllc). Technical specification 38.824 (March 2019). Version 2.0.1.2.
7. Yaacoub, E., & Alouini, M.-S. (2020). A key 6g challenge and opportunity-connecting the base of the pyramid: A survey on rural connectivity. *Proceedings of the IEEE, 108*(4), 533–582.
8. Akhtar, M. W., Hassan, S., Ghaffar, R., Jung, H., Garg, S., & Hossain, M. S. (2020). The shift to 6g communications: Vision and requirements.
9. Pocovi, G., Pedersen, K., & Mogensen, P. (2018). Joint link adaptation and scheduling for 5g ultra-reliable low-latency communications. *IEEE Access*, *6*, 28912 - 28922.
10. Popovski, P., Trillingsgaard, K., Simeone, O., & Durisi, G. (2018). 5g wireless network slicing for embb, urllc, and mmtc: A communication-theoretic view. *IEEE Access, 6*, 55765–55779.
11. 3GPP: Technical specification group services and system aspects; release 15 description. Technical report 21.915 (March 2019). Version 1.1.0.
12. 3GPP: 5g; nr; physical channels and modulation. Technical specification 38.211 release 16 (July 2020). Version 16.2.0.
13. 3GPP: 5g; nr; physical layer procedures for control. Technical specification 38.213 release 15 (October 2019). Version 15.7.0.
14. Al-Ali, M., Yaacoub, E., & Mohamed, A. (2020). Dynamic resource allocation of embb-urllc traffic in 5g new radio, pp. 1–6.
15. Panno, D., & Riolo, S. (2020). An enhanced joint scheduling scheme for GBR and non-GBR services in 5G RAN. *Wireless Networks, 26*, 3033–3052.
16. Akhtar, T., Tselios, C., & Politis, I. (2021). Radio resource management: Approaches and implementations from 4G to 5G and beyond. *Wireless Networks, 27*, 693–734.
17. Abdel Hakeem, S., Hady, A., & Kim, H. (2020). 5G-V2X: Standardization, architecture, use cases, network-slicing, and edge-computing. *Wireless Networks, 26*, 6015–6041.
18. Mouawad, N., Naja, R., & Tohme, S. (2020). Inter-slice handover management in a V2X slicing environment using bargaining games. *Wireless Networks, 26*, 3883–3903.
19. Zhang, R., Ning, L., Li, M., Wang, C., Li, W., & Wang, Y. (2021). Feature extraction of trajectories for mobility modeling in 5G NB-IoT networks. *Wireless Networks*, *27*, 1–13.
20. Anand, A., Veciana, G., & Shakkottai, S. (2020). Joint scheduling of urllc and embb traffic in 5g wireless networks. *IEEE/ACM Transactions on Networking*, *28*(2), 477–490.
21. Pradhan, A., & Das, S. (2020). Joint preference metric for efficient resource allocation in co-existence of eMBB and URLLC.
22. Zhang, X., Guo, X., & Zhang, H. (2021). Rb allocation scheme for embb and urllc coexistence in 5g and beyond. *Wireless Communications and Mobile Computing, 2021*, 1–7.
23. Manzoor, A., Kazmi, S. M., Pandey, S., & Hong, C. S. (2020). Contract-based scheduling of urllc packets in incumbent embb traffic.
24. Al-Senwi, M., Tran, N., Bennis, M., Pandey, S., Bairagi, A., & Hong, C. S. (2021). Intelligent resource slicing for embb and urllc coexistence in 5g and beyond: A deep reinforcement learning based approach. *IEEE Transactions on Wireless Communications*, *20*(7), 4585–4600.
25. Prathyusha, Y., & Sheu, T.-L. (2022). Coordinated resource allocations for embb and urllc in 5g communication networks. *IEEE Transactions on Vehicular Technology, 71*, 8717–8728.
26. Zhang, W., Derakhshani, M., & Lambotharan, S. (2020). Stochastic optimization of urllc-embb joint scheduling with queuing mechanism. *IEEE Wireless Communications Letters*, *10*(4), 844–848.
27. Almekhlafi, M., Arfaoui, M. A., Assi, C., & Ghrayeb, A. (2022). Superposition-based urllc traffic scheduling in 5g and beyond wireless networks. *IEEE Transactions on Communications, 70*, 1–1.
28. Yin, H., Zhang, L., & Roy, S. (2020). Multiplexing urllc traffic within embb services in 5g nr: Fair scheduling. *IEEE Transactions on Communications, 69*, 1080–1093.
29. Ferdosian, N., Skaperas, S., Chorti, A., & Mamatas, L. (2021). Conflict-aware multi-numerology radio resource allocation for heterogeneous services, pp. 1–6.
30. Darabi, M., Jamali, V., Lampe, L., & Schober, R. (2022). Hybrid puncturing and superposition scheme for joint scheduling of urllc and embb traffic. *IEEE Communications Letters, 26*, 1–1.
31. Afroz, F., Sandrasegaran, K., & Ghosal, P. (2015). Performance analysis of pf, m-lwdf and exp/pf packet scheduling algorithms in

3gpp lte downlink. In *2014 Australasian Telecommunication Networks and Applications Conference, ATNAC 2014*, pp. 87–92.

32. Yaacoub, E., & Dawy, Z. (2012). Resource allocation in uplink OFDMA wireless systems: Optimal solutions and practical implementations.
33. Musleh, S., Ismail, M., & Nordin, R. (2015). Effect of average-throughput window size on proportional fair scheduling for radio resources in lte-a networks. *Journal of Theoretical and Applied Information Technology, 80*, 179–183.
34. Andrews, M., Kumaran, K., Ramanan, K., Stolyar, A., Whiting, P., & Vijayakumar, R. (2001). providing quality of service over a shared wireless link. *IEEE Communications Magazine, 39*, 150–154.
35. Al-Senwi, M., & Hong, C. S. (2018). Resource scheduling of urllc/embb traffics in 5g new radio: A punctured scheduling approach.
36. Pandey, S., Al-Senwi, M., Tun, Y. K., & Hong, C. S. (2019). A downlink resource scheduling strategy for urllc traffic.
37. Al-Senwi, M., Pandey, S., Tun, Y. K., Kim, K., & Hong, C. S. (2019). A chance constrained based formulation for dynamic multiplexing of embb-urllc traffics in 5g new radio, pp. 108–113.
38. Sutton, G., Zeng, J., Liu, R., Ni, W., Nguyen, D., Jayawickrama, B., Huang, X., Abolhasan, M., Zhang, Z., Dutkiewicz, E., & Lv, T. (2019). Enabling technologies for ultra-reliable and low latency communications: From phy and mac layer perspectives. *IEEE Communications Surveys & Tutorials*, *21*(3), 2488–2524.
39. Yilmaz, O. N. C. (2016). Ultra-reliable and low-latency 5g communication, pp. 1–2.
40. Miao, G., Zander, J., Sung, K. W., & Slimane, S. B. (2016). Fundamentals of mobile data networks, pp. 1–304.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Muhammed Al-Ali received his bachelor's degree in computer engineering in 2013 from the University of Basrah in Basra, Iraq. He obtained his master's degree in computing from Qatar University in January 2021. He is currently working toward his Ph.D. degree at the Department of Computer Science and Engineering, Qatar University. His research interests include wireless communication, radio resource management, machine learning and cybersecurity.

Elias Yaacoub received the Bachelor of Engineering degree in Electrical Engineering from the Lebanese University in 2002, the Master of Engineering degree in Computer and Communications Engineering from the American University of Beirut (AUB) in 2005, and the Ph.D. degree in Electrical and Computer Engineering from AUB in 2010. He worked as a Research Assistant in the American University of Beirut from 2004 to 2005, and in the Munich University of Technology in Spring 2005. From 2005 to 2007, he worked as a Telecommunications Engineer with Dar Al-Handasah, Shair and Partners. From November 2010 till December 2014, he worked as a Research Scientist / R &D Expert at the Qatar Mobility Innovations Center (QMIC), where he led the Broadband Wireless Access Technology Team. Afterwards, he joined Strategic Decisions Group (SDG) where he worked as a Consultant till February 2016. Then he joined the Arab Open University (AOU) as an Associate Professor and Coordinator of the MSc Program in Information Security and Forensics. Between February 2018 and August 2019, he worked as an independent researcher/consultant, and he was also affiliated with AUB as a part-time faculty member. He joined Qatar University as an Associate Professor at the Computer Science and Engineering Department since August 2019. His research interests include Wireless Communications, Resource Allocation in Wireless Networks, Intercell Interference Mitigation Techniques, Antenna Theory, Sensor Networks, and Physical Layer Security.