Joint eMBB-URLLC Resource Allocation Based on Reliability Requirements of Users

Jianxiong Liu[1(✉)], Xiang Yu[1], Bei Liu[2], Xin Su[3], and Xibin Xu[3]

[1] School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing, China
S210131132@stu.cqupt.edu.cn

[2] Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing, China
liubei@tsinghua.edu.cn

[3] Department of Electronic Engineering, Tsinghua University, Beijing, China
{suxin,xuxb}@tsinghua.edu.cn

Abstract. The emerging fifth generation of mobile communication (5G) systems is designed to cater to heterogeneous services. The enhanced mobile broadband (eMBB) services and ultra-reliable and low-latency communication (URLLC) services are two of the core services. These two types of heterogeneous services have different requirements for network throughput and delay. This paper studies the joint resource allocation of eMBB and URLLC services in the co-existence scenario to maximize the reliability of eMBB services and average data rate. We propose a heuristic eMBB resource allocation scheme. In the URLLC resource allocation phase, we employ the non-orthogonal multiple access (NOMA) superposition techniques to multiplex the transmitted eMBB services and use one-to-one matching theory to evaluate optimal eMBB and URLLC user pairs to protect low-rate, high service requirement of eMBB users. The simulation results prove our method scheme achieves better reliability and average data rate of eMBB services.

Keywords: eMBB/URLLC coexistence · reliability · resource allocation · superposed transmission

1 Introduction

The rapid development of wireless communication has brought a variety of service demands. High data rate video streaming, Telematics, and Virtual Reality (VR) are among the key applications. However, throughput, delay requirements, and reliability are different for these services. eMBB services and URLLC services are two key services in 5G networks. eMBB services need to meet high data rates [1] while URLLC services need to meet low delay and high reliability [2]. In wireless communication systems, time is divided into time slots, and time slot

F. Gao et al. (Eds.): Chinacom 2023, LNICST 590, pp. 300–314, 2024.
https://doi.org/10.1007/978-3-031-67162-3_19

is further divided into mini-slots. To support the coexistence of services in the heterogeneous network, The Third Generation Partnership (3GPP) proposes the superposition/punch mechanism. eMBB services being transmitted in the punch technique are preempted by URLLC services arriving in each mini-slot, resulting in significant degradation of eMBB user throughput. Superposition technology in which eMBB services and URLLC services share time-frequency resources [3,4] poses the problem of superposition interference. In general, the punch technology brings more serious rate loss but no inter-user interference problems, while the superposition technology causes less rate loss but brings interference problems that require the use of successive interference cancellation (SIC) technology to decode at the receiving end, which brings additional delay costs. The trade-off between superposition and punch technology has become a difficult problem for resource scheduling. Meanwhile, different eMBB services have different minimum rate requirements. For example, a high data rate is required in a live webcast, if URLLC service arrives and grabs resources causing excessive loss of eMBB service rate, the service will be stopped so that the reliability of eMBB service is significantly reduced. Therefore, for the smooth playback of live broadcast, eMBB service needs to meet the minimum rate threshold, so that the image quality in a live webcast is reduced from high definition (HD) to standard definition (SD) to maintain the normal operation of the service and thus improve the reliability of eMBB service. Based on these considerations, we recommend a scheme for coordinated scheduling to improve the reliability of eMBB services.

Many scholars have studied the resource multiplexing problem in the coexistence scenario of eMBB services and URLLC services. The authors of [5] propose a matching strategy between eMBB and URLLC based on spatial multiplexing to maximize the sum rate of eMBB users. In [6], the authors proposed a non-orthogonal multiple access method to maximize the downlink system throughput. In [7] the authors propose to serve eMBB and URLLC services over the unlicensed spectrum and propose a preemptive approach. The authors of [8] propose a dynamic multiplexing approach based on deep reinforcement learning to mitigate the adverse effects due to URLLC preemption. The authors of [9] proposed a linear model, convex model, and threshold model for rate loss due to resource preemption of eMBB services by URLLC services and maximized the rate of eMBB services. However, these works [5–9] do not focus on the fact that different eMBB services are also required to ensure their reliability. The authors of [10] propose a risk-sensitive eMBB and URLLC resource multiplexing-based approach to protect eMBB users with low data rates. Although this work considers the reliability of eMBB services to some extent, its time-slot-based scheduling scheme is not suitable for URLLC services. Moreover, in [11], the authors proposed a punching scheme to minimize the decoding error rate of URLLC services based on ensuring the throughput requirements of eMBB users. The above-mentioned works either do not consider the reliability of eMBB services, use time slots as scheduling units that are not suitable for URLLC services, or use only punch scheduling schemes without weighing the choice of superposition and punch techniques. Therefore, in this paper, we propose a scheduling scheme with a trade-off between superposition

and punch techniques, aiming to improve the reliability and average data rate of eMBB users The main contributions of this paper are:

- In eMBB resource allocation, we propose a heuristic algorithm to improve the reliability and average data rate of eMBB users.
- For the resource allocation scheme of URLLC users, we construct preference lists aimed at improving the reliability and average data rate of eMBB users to reach a stable one-to-one matching pair based on matching theory. For paired users, we use NOMA technology for superposition transmission. When URLLC users cannot be paired, we use punching technology.
- Finally, we simulate the proposed approach and compare the average data rate and the reliability of eMBB users with the coordinated resource allocation [12] and the puncturing technique under the same conditions. Simulation results show that the proposed scheme achieves higher eMBB reliability and higher average data rate.

The rest of this paper is presented below. Section 2 outlines the background of dynamic multiplexing of eMBB and URLLC services. The solution is presented in Sect. 3. The simulation results are given in Sect. 4. Finally, we summarize in Sect. 5.

2 Model and Problem Building

The wireless communication system includes eMBB and URLLC users as well as a base station (BS), which is shown in Fig. 1. The system model contains eMBB user set $E = \{1, 2 \ldots e\}$ and URLLC user set $L = \{1, 2, \ldots, l\}$. The time slots denoted by $M = \{1, 2, \ldots, m\}$, RB set $R = \{1, 2, \ldots, r\}$ and the bandwidth of each RB is B. The length of time slot is τ. Each slot is divided into several mini-slots, denoted by $N = \{1, 2, \ldots, n\}$, and the length of mini-slot is ω. We assume that P is the total transmitted power of RB.

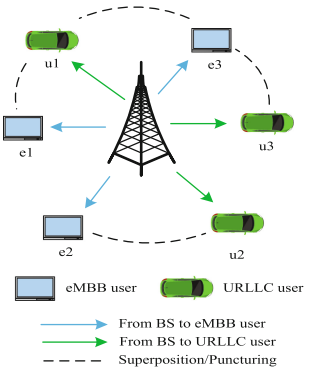

Fig. 1. System model of eMBB and URLLC services in downlink scenario.

2.1 URLLC Services

URLLC users arrive during the eMBB user transmission. To achieve the strict QoS for URLLC services, they need to be served in the next mini-slot immediately. Due to the short URLLC packet size, it is not feasible to use Shannon's capacity formulation, so we use finite block-length coding. User $l \in L$ of URLLC is given the decoding error probability ε_l, and the rate that can be achieved under the finite block length G_l is approximately as follows [13,14]:

$$R_{l,r}^{n,m}\left(\gamma_{l,r}^{n,m}\right) = B\log_2\left(1+\gamma_{l,r}^{n,m}\right) - \sqrt{\frac{V_l}{G_l}}\frac{Q^{-1}\left(\varepsilon_l\right)}{\ln 2} \tag{1}$$

where $\gamma_{l,r}^{n,m}$ indicates the signal-to-noise ratio (SNR) of URLLC user l at mini-slot n of time slot m on RB r. The SNR has different expressions in different superposition techniques. Let $\hat{\gamma}_{l,r}^{n,m}$ and $\bar{\gamma}_{e,r}^{n,m}$ denote SNR of superposed and punch transmission. $Q^{-1}(\cdot)$ is the inverse of $Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-m^2/2}\, dt$, and the channel dispersion $V_l = \left(1 - \left(1+\gamma_{l,r}^{n,m}\right)^{-2}\right)$ based on [15]. As can be seen from Fig. 2, the URLLC service is transmitted overlapped with the eMBB service being transmitted, which brings data rate loss to the overlapped eMBB service.

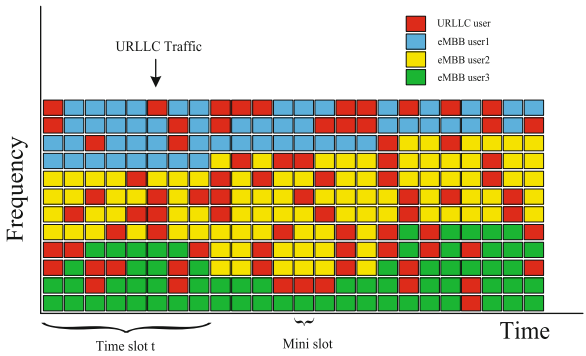

Fig. 2. Superposition/piercing scenario of eMBB and URLLC services.

2.2 NOMA Superposition

We use the NOMA technique for superposition transmission. Considering two paired eMBB user e and URLLC user l, the BS transmits the following superposed signal:

$$S = P_e S_e + P_l S_l \tag{2}$$

The signals of user e and l are S_e and S_l respectively. And P_e and P_l are their corresponding allocated power. We assume that $P = P_e + P_l$ is satisfied.

We assume that h_e and N_e represent eMBB's channel gain and noise, while h_l and N_l represent URLLC's channel gain and noise. The signal received from the user is as follows:

$$Y_e = h_e \left(P_e S_e + P_l S_l\right) + N_e \tag{3}$$

$$Y_l = h_l \left(P_e S_e + P_l S_l\right) + N_l \tag{4}$$

For the matched eMBB user and URLLC user, the successive interference cancellation technique (SIC) is performed at the user with high channel gain. The SNR's of eMBB user and URLLC user are: $\hat{\gamma}_{e,r}^{n,m} = \frac{|h_e|^2 P_e}{N_e}, \hat{\gamma}_{l,r}^{n,m} = \frac{|h_l|^2 P_l}{|h_l|^2 P_e + N_l}$ and vice versa. If the NOMA technology is not available for URLLC users, then punch technology will be used. So the SNR of an eMBB user and a URLLC user are $\bar{\gamma}_{e,r}^{n,m} = 0$ and $\bar{\gamma}_{l,r}^{n,m} = \frac{|h_l|^2 P_l}{N_l}$. If eMBB users do not superpose pairs, the SNR is $\tilde{\gamma}_{e,r}^{n,m} = \frac{|h_e|^2 P_e}{N_e}$.

When URLLC user l needs multiple RBs to meet its QoS, the implementation rate is as follows:

$$R_l^{n,m} = \sum_{r \in R} \left\{ \begin{array}{l} \alpha_{e,l,r}^{n,m} * R_{l,r}^{n,m} \left(\hat{\gamma}_{l,r}^{n,m}\right) \\ + \beta_{e,l,r}^{n,m} * R_{l,r}^{n,m} \left(\bar{\gamma}_{l,r}^{n,m}\right) \end{array} \right\} \tag{5}$$

where α is the distribution vector of RB representing superposition transmission, as follows:

$$\alpha_{e,l,r}^{n,m} = \begin{cases} 1, \text{ if eMBB user } e \text{ and URLLC user } l \\ \quad \text{are transmitted superimposed,} \\ 0, \text{ Else.} \end{cases} \tag{6}$$

β is an assigned vector of an RB indicating the punch transmission of the eMBB and URLLC user expressed as follows:

$$\beta_{e,l,r}^{n,m} = \begin{cases} 1, \text{ if eMBB user } e \text{ and URLLC user } l \\ \quad \text{punch the transmission,} \\ 0, \text{ Else.} \end{cases} \tag{7}$$

where $\bar{\alpha}, \bar{\beta}$ are complements. This payload $S_l^{n,m}$ of the URLLC user l must be delivered in the prescribed time limit ω. The delay requirements are guaranteed by the following equation:

$$v_l^{n,m} S_l^{n,m} \leq \omega R_l^{n,m}, \forall l \in L, m \in M, n \in N \tag{8}$$

where $v_l^{n,m}$ is the distribution vector for l, as follows:

$$v_l^{n,m} = \begin{cases} 1, \text{ if } l \text{ is accepted at} \\ \quad n \text{ of } m, \\ 0, \text{ Else.} \end{cases} \tag{9}$$

Equation (8) ensures that for the duration of a mini-slot, the achieved rate of the URLLC user is greater than the minimum rate required by the user to

achieve payload transmission. As shown in (5), an URLLC user needs to allocate multiple RBs to satisfy payload transmission.

The constraint of the URLLC service is that all URLLC users must complete the service within the duration of the mini-slot. It is expressed as:

$$P\left(\sum_{l\in L} v_l^{n,m} < K\right) \leq (1-\varepsilon), \forall n \in N, m \in M \tag{10}$$

Assume that the number of URLLC user arrivals obeying a Poisson distribution with mean λ is K and the reliability probability of URLLC users is $\epsilon = 0.999$. Then the probability of ensuring that the number of served URLLC users is less than the number of arriving URLLC users in (10) must be less than 0.001. Thus the reliability of URLLC is satisfied.

The rate of an eMBB user e at RB $r \in R$ is expressed as follows:

$$R_{e,r}^{n,m}\left(\gamma_{e,r}^{n,m}\right) = B\log_2\left(1+\gamma_{e,r}^{n,m}\right) \tag{11}$$

The achieved rate of eMBB user $e \in E$ is expressed as follows:

$$R_e^{n,m} = \sum_{r\in R} x_{e,r}^m \left\{ \begin{array}{l} \alpha_{e,l,r}^{n,m} * R_{e,l}^{n,m}\left(\hat{\gamma}_{e,r}^{n,m}\right) + \\ \beta_{e,l,r}^{n,m} * R_{e,r}^{n,m}\left(\bar{\gamma}_{e,r}^{n,m}\right) + \\ \bar{\alpha}_{e,l,r}^{n,m} * \bar{\beta}_{e,l,r}^{n,m} * R_{e,r}^{n,m}\left(\tilde{\gamma}_{e,r}^{n,m}\right) \end{array} \right\} \tag{12}$$

We assume that x represents the frequency resource allocation vector for e, expressed as:

$$x_{e,r}^m = \begin{cases} 1, \text{ if RB } r \text{ is assigned to} \\ \quad e, \\ 0, \text{ Else.} \end{cases} \tag{13}$$

The actual data rate achieved by eMBB user is expressed as follows:

$$R_{e,act}^m = \sum_{n\in N} R_e^{n,m} \tag{14}$$

Let Q_e^m denote the reliability index of eMBB user e in slot m which can be expressed as:

$$Q_e^m = \begin{cases} 1, \text{ if } R_{e,act}^m \geq R_{e,th}^m, \\ 0, \text{ Otherwise.} \end{cases} \tag{15}$$

where $R_{e,th}^m$ represents the minimum implementation rate that eMBB user $e \in E$ needs to achieve in slot $m \in M$.

2.3 Problem Formulation

This work consists of resource scheduling between eMBB and URLLC services. eMBB resources are allocated to improve the reliability of eMBB. While URLLC

services superposition/punch the eMBB services being transmitted to meet their QoS and select the optimized URLLC-eMBB pair without affecting the reliability of eMBB and maximize the average speed of eMBB users. The formula for time slot optimization is as follows:

$$\max_{x} \sum_{e\in E} Q_e^m + \max_{\alpha,\beta,\psi} \sum_{n\in N}\sum_{r\in R}\sum_{l\in L}\sum_{e\in E}\left(\alpha_{e,l,r}^{n,m}+\beta_{e,l,r}^{n,m}\right)\left(R_e^{n,m}\right) \tag{16}$$

$$v_l^{n,m} S_l^{n,m} \le \omega R_l^{n,m}, \forall l \in L, n \in N, m \in M, \tag{16a}$$

$$P\left(\sum_{l\in L} v_l^{n,m} < K\right) \le (1-\varepsilon), \forall n \in N, m \in M, \tag{16b}$$

$$\sum_{e\in E} x_{e,r}^m \le 1, \forall r \in R, m \in M, \tag{16c}$$

$$\sum_{r\in R}\left(\alpha_{e,l,r}^{n,m}+\beta_{e,l,r}^{n,m}\right) \ge 1, \forall e \in E, l \in L, n \in N, m \in M, \tag{16d}$$

$$\sum_{e\in E}\sum_{l\in L}\left(\alpha_{e,l,r}^{n,m}+\beta_{e,l,r}^{n,m}\right) \le 1, \forall n \in N, m \in M, \tag{16e}$$

$$\sum_{r\in R} x_{e,r}^m \ge C_e^m, \forall m \in M, \forall e \in E \tag{16f}$$

$$\sum_{r\in R}\sum_{e\in E} x_{e,r}^m \le |R|, \forall m \in M, \tag{16g}$$

$$\sum_{r\in R}\sum_{e\in E}\sum_{l\in L}\left(\alpha_{e,l,r}^{n,m}+\beta_{e,l,r}^{n,m}\right) \le |R|, \forall n \in N, m \in M, \tag{16h}$$

$$x_{e,r}^m, \alpha_{e,l,r}^{n,m}, \beta_{e,l,r}^{n,m}, v_l^{n,m} \in (0,1), \\ \forall e \in E, l \in L, r \in R, n \in N, m \in M.. \tag{16i}$$

Delay and reliability constraints for URLLC users are ensured in (16a) and (16b). Equation (16c) ensure that each RB can only be owned by one eMBB user. Constraint (16d) guarantees that the served URLLC users are assigned to at least one RB. In (16e), each RB is owned by one URLLC user. Constraint (16f) ensures that each eMBB user is assigned at least the number of RBs required to satisfy its minimum rate requirement before its resources are seized by URLLC users, where C_e^m is the minimum number of RB for eMBB user $e \in E$ to meet its minimum rate requirement. Constraint (16g) indicates that the total number of RBs allocated to all eMBB users cannot exceed the system resources. In (16h), the total number of RBs allocated to all URLLC users cannot exceed the system resources. Finally (16i) states that each vector element is binary.

3 Solution Approach

For the problem in Eq. (16), we use an alternate optimization method to divide the original problem into two subproblems. There are the time-slot-based eMBB resource allocation problem and the mini-slot-based power allocation and pairing problem for URLLC users and eMBB users. The first subproblem aims to maximize the reliability of eMBB services, and the second subproblem aims to maximize the average rate of eMBB users without affecting the reliability of eMBB services.

3.1 eMBB Resource Allocation

eMBB resource allocation aims to maximize the reliability of eMBB services. Therefore, we propose a heuristic algorithm in Algorithm 1 to allocate eMBB resources at the boundaries of time slots to solve the first subproblem.

- Step1: Each eMBB user has different $R^m_{e,th}$ in the slot m, where C_B represents the total number of remaining available RB and $R^m_{e,RB}$ represents the rate of an RB achieved by eMBB user $e \in E$ in slot m.
- Step2: Sort S_e in ascending order of the size of $R^m_{e,th}$, where S_e represents the eMBB user set.
- Step3: If there are remaining RBs in the system, the eMBB users are sorted in descending order according to the channel gain $|h_e|^2$.

When the frequency resources of the wireless communication system are insufficient, the resource allocation scheme maximizes the number of eMBB users meeting the minimum rate requirement. Thus, the reliability of eMBB service is improved. If the system frequency resources are sufficient, the minimum rate requirements for all eMBB users can be guaranteed, and the average data rate can be further improved.

In Algorithm 1, we let β as the number of resource block which are given to eMBB users at slot m in the first step assignment, where $\beta \leq |R|$. We assume that $(|R| - \beta)$ indicates the remaining number of resource blocks assigned to eMBB users. The required complexity for each RB assignment is $O(1)$. The time complexity required is $O(|R| - \beta)$. So the total time complexity of Algorithm 1 is $O(|M|\,|R|)$.

3.2 URLLC Allocation

Frequency resource pairing includes match and the allocation of power required by URLLC users. URLLC users can choose the superposition or punch techstanique to retain in the RBs occupied by eMBB users when they arrive. Before the overlay transmission, each URLLC user needs to be assigned the power required to satisfy its QoS. We use one-to-one matching theory and NOMA superposition techniques to solve the second subproblem. We assume that $S^{n,m}_l$ is the packet size of a URLLC user $l \in L$ in mini-slot n of time slot m. When

Algorithm 1 : eMBB Resource Allocation.

Initialization: $C_B = N, C_e^m = 0, R_{e,RB}^m = 0, k = 0, i = 0, j = 0, m = 0$
while $m \leq |M|$ **do**
  Sort eMBB users set S_e according to the $R_{e,th}^m$ value
  the ascending order.
  Get $R_{e,th}^m$ for each eMBB user at time slot boundary.
  for $k = 1 : |E|$ **do**
    $C_e^m = \left\lceil \frac{R_{e,th}^m}{R_{e,RB}^m} \right\rceil$
    if $C_B > Ce^m$ **then**
      $E(k) \rightarrow nRBs = C_e^m$
      Update $C_B = C_B - C_e^m$
    else
      $E(k) \rightarrow nRBs = C_B$
    end if
  end for
  Update C_B and resort S_e according to the $|he|^2$ in
  the descending order.
  for $i = 1 : C_B$ **do**
    $j = i \bmod |E|$
    $E(j) \rightarrow RBs = E(j) \rightarrow RBs + 1$
  end for
  Update $m = m + 1$
end while

the URLLC user uses the superposition technique, we calculate the required power $p_{l,e,req}^{n,m}$ to satisfy the QoS constraint for this URLLC user. $\psi_l^{n,m}$ denotes an eMBB user $e \in E$ paired by an URLLC user $l \in L$.

In this paper, no more than two users are superposed using the NOMA technique.

Definition 1: Matches $\varphi^{n,m}$ defined as a function of the sets $|E| \cup |L|$ through $|L| \cup |E|$, and needs to satisfy the following constraints: 1) $\varphi^{n,m}(e) \subseteq L, \forall e \in E$
2) $\varphi^{n,m}(l) \subseteq E, \forall l \in L$
3) $|\varphi^{n,m}(e)| \leq 1, \forall e \in E$
4) $|\varphi^{n,m}(l)| \leq 1, \forall l \in L$

Definition 2: We initialize the preference list of URLLC users according to the following SIC feasibility conditions:

$$\psi_l^{n,m} = e, \forall e \in E, p_{l,e,req}^{n,m} < \frac{P}{2} \tag{17}$$

$$\psi_l^{n,m} = e, \forall e \in E, p_{l,e,req}^{n,m} > \frac{P}{2}, |h_l|^2 < |h_e|^2 \tag{18}$$

Pairing in the power domain using NOMA requires following the SIC decoding order. On the one hand, SIC can only be performed for users with strong channel gain to eliminate interference. On the other hand, since weak users need

to be allocated high power, in (17), a URLLC user can be paired with any eMBB user if the power required by that user is less than $\frac{P}{2}$. In (18), however, the required power of the URLLC user is higher than $\frac{P}{2}$.

Let e_1 and e_2 as two eMBB users. Under the condition that $|h_e|^2$ and $R_{e,th}^m$ is the same, if $R_{e\_act}^{n,m}$ of e_1 is lower than $R_{e\_act}^{n,m}$ of e_2, where $R_{e\_act}^{n,m}$ represents the real rate realized by eMBB user in mini-slot n of slot t. URLLC users prefer e_1 to e_2, which makes it easier for e_1 to achieve the minimum rate requirement. This is because compared with punching technology, NOMA superposition achieves a higher rate, which is more conducive for eMBB users to achieve their minimum rate requirement. Similarly, when all other conditions are equal, if the $R_{e,th}^m$ or $|h_e|^2$ of e_1 is higher than e_2, URLLC user prefers e_1. The preference list of user URLLC is as follows:

$$e_1 \prec_l e_2, \{e_1, e_2\} \in E, \tag{19}$$

According to the following conditions:

$$\max\left(\frac{|h_e|^2 R_{e,th}^m}{R_{e\_act}^{n,m}}\right) \tag{20}$$

Definition 3: eMBB users prefer to be matched with the URLLC user which maximizes their rate. Thus the preference list of eMBB users is constructed as shown below:

$$l_1 \prec_e l_2, \{l_1, l_2\} \in L, \tag{21}$$

Let l_1, l_2 denote two URLLC users. We construct the preference list based on the following objective function:

$$\max\left(R_e^{n,m}\right) \tag{22}$$

After constructing a preference matching list for both parties, we propose a algorithm to output stable matching pairs.

Theorem 1: The algorithm can output stable matching pairs.

Proof:

Definition 4: We say that the match $\varphi^{n,m}$ has a blocking pair if there is a pair of (e, l) in the match and $e \in E, l \in L, e \notin \varphi^{n,m}(l), l \notin \varphi^{n,m}(e)$ and the following two propositions are also satisfied:

1) $u \prec_e \varphi^{n,m}(e)$
2) $e \prec_l \varphi^{n,m}(l)$

Definition 5: We call a match stable if it has no blocking pairs.

Lemma: The match algorithm can output of this algorithm is stable.

Proof: In Algorithm 2, assuming that at ith iteration in mini-slot n, the updated preference lists of URLLC user l and eMBB user e from the $i-1$ iteration are

defined by $\rho_e^{n,m}$ and $\rho_l^{n,m}$, respectively. When an eMBB user e rejects a URLLC user l at one iteration, the eMBB user e is removed from the preference list of l, and vice versa. The successfully matched URLLC users in the same iteration will be retained for the next iteration. Additionally, we adopt the Deferred Acceptance Algorithm (Gale-Shapley, or GS) to generate suitable eMBB-URLLC pairings, where it is proven that the matching produced by this algorithm is stable [16]. Therefore, this algorithm can output stable matchings and converge within a finite number of iterations.

Algorithm 2 : eMBB and URLLC user matching algorithm.

Initialization : $\rho_e^{n,m}$, $\rho_l^{n,m}$, $\varphi^{n,m}(e)$, $\varphi^{n,m}(l) = \emptyset$, $\forall e \in E, l \in L$
Let i be the iteration variable and have an initial value of 0
Let $\widehat{l}$ be the URLLC user matched by the current iteration of the eMBB user
Output stability matching
repeat
    $i = i + 1$
    for $l \in L$ chose $\rho_l^{n,m}$ **do**
        if $l \prec_e \varphi^{n,m}(e)$ **then**
            $\varphi^{n,m}(e) \leftarrow \varphi^{n,m}(e)/\widehat{n}$
            $\varphi^{n,m}(e) \leftarrow l$
            $\rho_e^{n,m} \leftarrow \rho_e^{n,m}/\widehat{n}$
            $\rho_{\widehat{l}}^{n,m} \leftarrow \rho_{\widehat{l}}^{n,m}/e$
        else
            $\rho_e^{n,m} \leftarrow \rho_e^{n,m}/l$
            $\rho_l^{n,m} \leftarrow \rho_l^{n,m}/e$
        end if
    end for
until $\varphi^{n,m}$ was successfully matched in the previous iteration
Generate the corresponding $\varphi^{n,m}(l), l \in L$ from $\varphi^{n,m}(e), e \in E$
for $l \in L$ **do**
    if $\varphi^{n,m}(l) = \emptyset$ **then**
        We use puncture technology for transmission
    else
        The corresponding superposition technology is adopted for transmission according to $\varphi^{n,m}(l)$
    end if
end for

In Algorithm 2, the time complexity required to construct the preference list for URLLC users using a standard sorting algorithm is $O(|E| \log |E|)$, and thus, the time complexity is $O(|L||E| \log |E|)$. Similarly, the time complexity required for all eMBB users is $O(|E||L| \log |L|)$. Therefore, the time complexity for constructing preference lists for all eMBB users and URLLC packets is $O(|E||L| \log |E||L|)$. Since Algorithm 1 converges and terminates within at most $|E||L|$ iterations, the time complexity required for this step is $O(|E||L|)$. Consequently, Algorithm 1 has a polynomial time complexity of $O(|E||L| \log |E||L|)$.

4 Simulation

In this part, we simulated various indicators according to the actual operating environment to evaluate the performance of our scheme. We consider a wireless network that consists of one BS. The number of eMBB users is $|E| = 10$. URLLC traffic follows a Poisson distribution with an average of λ, and then arrives in mini time slots. Moreover, $R^m_{e,th}$ at each time slot m is a random number in the range of 16–22 Mbps, which is different for each eMBB user. We consider the total power of an RB to be $P = 30$ dBm, and the total number of system RBs is $|R| = 50$. We use $B = 120$ KHz as the bandwidth size of an RB. We assume the duration of the time slot is $\tau = 1$ ms and the duration of a mini-slot is $\omega = 0.125$ ms. Meanwhile, we set the Gaussian white noise power spectral density as $N_0 = 10^{-14.4}$ W/Hz, and the noise power of eMBB users is denoted as $N_e = BN_0$. In addition, based on the URLLC reliability constraint, we choose $\varepsilon = 0.999$ and the payload $S_l^{n,m} = 32$ bit for $l \in L$. In the following, we compare the proposed scheme with the other two schemes in different performance indicators under the same system model:

(1) The scheme in reference [12] (Coordinated): In reference [12], the author uses a scheme of combined superposition/punching techniques to maximize the minimum expected achieved rate (MEAR).
(2) Random Punch (Punch): Select the eMBB user to be punched evenly among all existing eMBB users. Due to the optimality of the random layout of the linear loss model, random punching is considered as a comparison scheme. This stems from the fact that if eMBB resources are punched evenly, the resources punched are proportional to the bandwidth allocated to each eMBB user.

Here are three performance evaluation parameters:

1) The average eMBB data rate is as follows:

$$\text{MeanRate} = \frac{1}{|E|\,|M|} \sum_{m=1}^{|M|} \sum_{e \in E} R^m_{e,act} \tag{23}$$

2) MEAR for eMBB users is represented as follows:

$$\text{MEAR} = \min \left(\frac{1}{|M|} \sum_{m=1}^{|M|} R^m_{e,act} \right), \forall e \in E \tag{24}$$

3) eMBB Reliability is evaluated as:

$$\text{eMBB Reliability} = \frac{\sum_{m \in M} \sum_{e \in E} Q^m_e}{|M|\,|E|} \tag{25}$$

In Fig. 3, it can be seen that as the λ increases, more eMBB services being transmitted have to allocate resources to URLLC users which cause the average

rate decreases. Compared to [12], our proposed algorithm improves in the average rate metric. This is because in [12], the algorithm allocates RBs to eMBB users with high rate losses in the previous time slot rather than to users with high channel conditions, thus reducing the average rate. Also Compared to the punching algorithm, the use of the superposition technique results in less rate loss for eMBB users, and thus the rate improvement is larger.

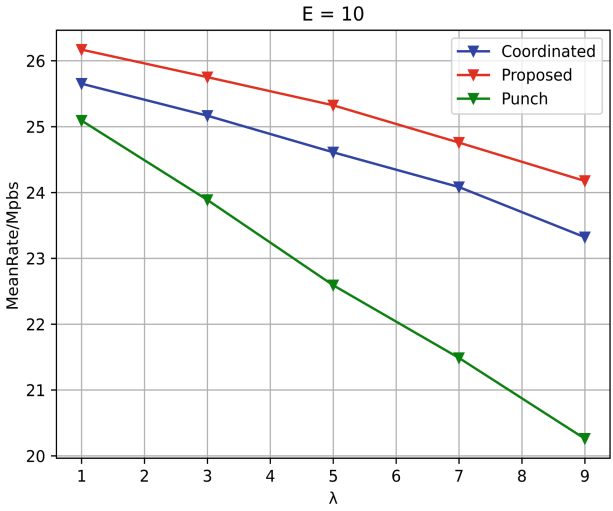

Fig. 3. λ versus MeanRate.

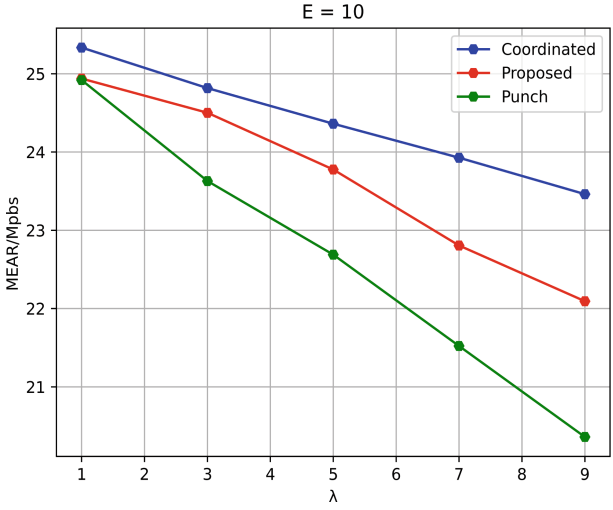

Fig. 4. λ versus MEAR.

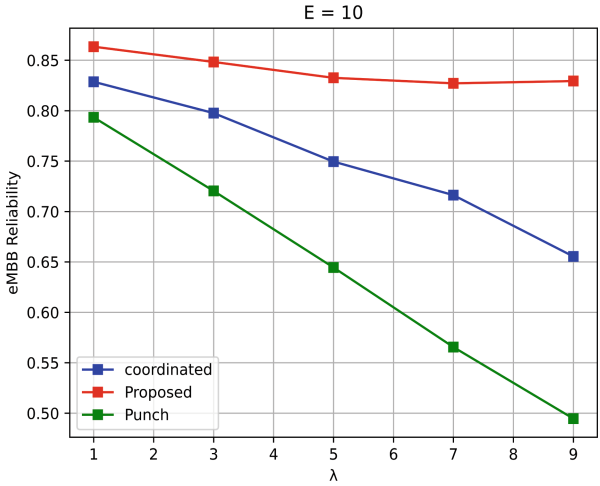

Fig. 5. λ versus eMBB Reliability.

Figure 4 illustrates the MEAR performance metric, where MEAR denotes the minimum expected average rate of eMBB users. The MEAR decreases as the number of URLLC user arrivals increases. Compared to [12], our proposed algorithm is lower in MEAR since the literature aims to maximize MEAR, while this work studies to improve the reliability and throughput of eMBB services while further increasing the throughput of eMBB services with sufficient resources.

From Fig. 5, we can see that the reliability of eMBB services decreases as the number of URLLC user arrivals increases. The reason for the superiority of our proposed algorithm over other algorithms is that we give priority to eMBB users with smaller minimum rate requirements when allocating eMBB resources, which increases the number of eMBB users satisfying minimum rate requirements and thus improves the reliability to some extent. However, the reliability requirements of eMBB are not considered in [12] and punching schemes.

5 Conclusion

In this article, we considered the conditions under which an eMBB service needs to meet the minimum rate requirement. We propose a resource allocation scheme to maximize the reliability and average data rate of the eMBB service. The scheme is divided into two parts, the time-slot based eMBB frequency resource allocation scheme and the mini-slot based URLLC resource allocation scheme. In the eMBB resource allocation scheme, we adopt a low complexity eMBB allocation algorithm to improve the reliability and average data rate of eMBB users. For URLLC resource allocation, we use matching theory to calculate suitable eMBB URLLC pairs, and adopt NOMA technology for superposition transmission, while for URLLC users who cannot be paired, we use punching technology.

It can be seen from the simulation results that the proposed scheme has a significant performance gain compared with the other two schemes for the indicators under consideration.

Acknowledgements. This work is supported by National Key R&D Program of China, No. 2020YFB1806702.

References

1. Alsenwi, M., Pandey, S.R., Tun, Y.K., Kim, K.T., Hong, C.S.: A chance constrained based formulation for dynamic multiplexing of eMBB-URLLC traffics in 5G new radio. In: Proceedings of International Conference on Information Networking, pp. 108–113 (2019)
2. Study on physical layer enhancements for NR ultra-reliable and low latency case (URLLC), document TR38.824, 3GPP (2019)
3. Physical channels and signals for 5G-NR, document TS 38.211, 3GPP (2018)
4. Downlink multiplexing of eMBB and uRLLC transmission, document R11700- 374, 3GPP TSG RAN WG1 NR Ad-Hoc Meeting, 3GPP (2017)
5. Chen, Q., Jiang, H., Yu, G.: Service oriented resource management in spatial reuse-based C-V2X networks. IEEE Wireless Commun. Lett. **9**(1), 91–94 (2020)
6. Shikuma, T., Yuda, Y., Higuchi, K.: NOMA-based optimal multiplexing method for downlink service channels to maximize integrated system throughput. In: 2019 IEEE 90th Vehicular Technology Conference (VTC2019-Fall), pp. 1–5 (2019)
7. Zaki-Hindi, A., Elayoubi, S., Chahed, T.: URLLC and eMBB coexistence in unlicensed spectrum: a preemptive approach. In: International Wireless Communications and Mobile Computing (IWCMC), pp. 229–234 (2020)
8. Huang, Y., Li, S., Li, C., Hou, Y.T., Lou, W.: A deep-reinforcement-learning-based approach to dynamic eMBB/URLLC multiplexing in 5G NR. IEEE Internet Things J. **7**(7), 6439–6456 (2020)
9. Anand, A., de Veciana, G., Shakkottai, S.: Joint scheduling of URLLC and eMBB traffic in 5G wireless networks. IEEE/ACM Trans. Netw. **28**(2), 477–490 (2020)
10. Alsenwi, M., Tran, N.H., Bennis, M., Kumar Bairagi, A., Hong, C.S.: eMBB-URLLC resource slicing: a risk-sensitive approach. IEEE Commun. Lett. **23**(4), 740–743 (2019)
11. Ning, W., Wang, Y., Liu, M., Chen, Y., Wang, X.: Mission-critical resource allocation with puncturing in industrial wireless networks under mixed services. IEEE Access **9**, 21870–21880 (2021)
12. Prathyusha, Y., Sheu, T.-L.: Coordinated resource allocations for eMBB and URLLC in 5G communication networks. IEEE Trans. Veh. Technol. **71**(8), 8717–8728 (2022). https://doi.org/10.1109/TVT.2022.3176018
13. Polyanskiy, Y., Poor, H.V., Verdu, S.: Channel coding rate in the finite blocklength regime. IEEE Trans. Inf. Theory **56**(5), 2307–2359 (2010)
14. Yang, W., Durisi, G., Koch, T., Polyanskiy, Y.: Quasi-static multipleantenna fading channels at finite blocklength. IEEE Trans. Inf. Theory **60**(7), 4232–4265 (2014)
15. Sun, X., Yan, S., Yang, N., Ding, Z., Shen, C., Zhong, Z.: Downlink NOMA transmission for low-latency short-packet communications. In: Proceedings of IEEE International Conference on Communications Workshops, pp. 1–6 (2018)
16. Gale, D., Shapley, L.S.: College admissions and the stability of marriage. Amer. Math. Monthly **69**(1), 9–15 (1962)