eMBB-URLLC Multiplexing: A Greedy Scheduling Strategy for URLLC Traffic with Multiple Delay Requirements

Can Li[1], Bei Liu[2], Xin Su[3], Xibin Xu[3]
[1]School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications Chongqing, China
[2]Beijing National Research Center for Information Science and Technology, Tsinghua University Beijing, China
[3]Department of electronic engineering, Tsinghua University Beijing, China
Corresponding Author: Can Li Email: s210101071@stu.cqupt.edu.cn

***Abstract*—The coexistence of enhanced mobile broadband (eMBB) and ultra-reliable low-latency communication (URLLC) is common in 5G networks. 5G services require eMBB users to achieve higher data rate, and URLLC users to meet high reliability and low latency requirements. How to utilize limited resources to maximize the throughput of eMBB users in the system while meeting URLLC latency requirements is a very meaningful issue. In this paper, we delve into the latency composition of URLLC packets and subsequently derive an expression to determine the number of mini-slots that URLLC packets can be queued. We propose a greedy scheduling algorithm based on queuing theory, which solves the complex scheduling problem of URLLC packets with different latency requirements. At each mini-slot, we dynamically schedule the URLLC packets that arrive using the proposed algorithm. We demonstrate the significant advantages of our algorithm compared to other algorithms through extensive simulations. Specifically, our algorithm significantly reduces the throughput loss of eMBB users, and also meets the high reliability requirements of URLLC in the case of high URLLC load.**

***Index Terms*—URLLC, eMBB, resource scheduling, different latency requirements**

I. Introduction

As a new generation of mobile communication technology, 5G mainly includes three scenarios: enhanced mobile bandwidth (eMBB), ultra-reliable and low-latency communication (URLLC), and massive machine type communication (mMTC) [1]. The coexistence and transmission of eMBB services and URLLC services has become a typical scenario. According to the 3GPP standard, the goal of eMBB service is to provide each user with a data rate exceeding 100Mb/s. URLLC service is a key in supporting 5G scenarios [2]–[4]. 3GPP summarizes the delay and reliability requirements of URLLC packets in different scenarios. Ensuring the high reliability and low latency requirements of URLLC while maximizing the data rate of eMBB users has become a crucial issue.

The coexistence of URLLC services and eMBB services has been extensively investigated. In [5], the authors investigated the joint scheduling problem of eMBB and URLLC and analyzed the linear, convex, and threshold models for eMBB rate loss due to superposition/puncturing. Authors in [6] proposed a risk-sensitive approach to transmit more URLLC packets to eMBB users with higher data rates, protecting eMBB users with lower data rates. In [7], the authors exploited the possible similarity between URLLC and eMBB symbols and selected the scheme that maximizes the number of similar symbols between the two services for resource punching multiplexing, which can effectively reduce the affected eMBB symbols and improve the used error correction code rate and spectral efficiency. In [8], a penalized successive upper bound minimization (PSUM) algorithm was used to schedule the eMBB, and an optimal transport model (TM) was used to schedule the URLLC, which maximizes the minimum expected achieved rate (MEAR) of the eMBB user equipment while satisfying the URLLC traffic requirements.

In the above works, the arriving URLLC packets were considered for immediate transmission at the current mini-slot, and if they cannot be transmitted, the URLLC packets will be discarded, which not only increased the probability of interruption of URLLC but also reduced the throughput of eMBB users. This situation can be avoided by using appropriate queuing mechanisms. In [9], [10], the authors introduced a queuing theory mechanism to queue URLLC packets into two mini-slots to reduce the throughput loss of eMBB users in the system. In [11], the authors introduced the M/G/1 queuing theory model and used "conditional value at risk (CVaR)" to evaluate whether the delay of URLLC data exceeds the threshold value, which maximized the eMBB data rate while satisfying the URLLC delay.

By incorporating a queueing theory model, the aforementioned study offers notable advancements in terms of both total eMBB user throughput and the reliability of URLLC. But queuing all URLLC packets into two mini-slots is not an optimal solution. We investigate the latency requirement of URLLC packets in a more fine-grained manner. We analyze the number of mini-slots that URLLC packets with different delay requirements can be queued. We create a queue. The length of the queue is the maximum number of mini-slots that URLLC packets in the system can be queued.

We use a greedy algorithm to dynamically maintain a queue. At each mini-slot, we use the proposed algorithm to determine the URLLC packets that need to be transmitted in the mini-slot based on the URLLC packets in the queue and

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

the arriving URLLC packets. Then, we update the queue. In this paper, by assigning the arriving URLLC data packets to the queue for transmission, small batch and multiple transmission of URLLC data packets is realized. Compared with previous work, our algorithm can effectively improve the system throughput and ensure the reliability requirements of URLLC users.

The main contributions of this paper are:

- We investigate the main delay composition of URLLC packets and derive a formula for mini-slots that URLLC data packets can be queued. We calculate the number of mini-slots that URLLC packets in different scenarios can be queued according to the formula.
- We create a queue with a length equal to the number of mini-slots that URLLC packets with the maximum latency requirement can be queued.
- We propose an algorithm to dynamically update the queue based on the arrived URLLC packets and determine the URLLC packets that need to be transmitted in the current mini-slot.

The rest of the paper is structured as follows. In Section II we introduce a system model. In Section III we distribute the arriving URLLC packets to multiple mini-slots for transmission. In Section IV we introduce the scheduling of URLLC. In Section V we present the simulation results. Finally, we give a conclusion in Section VI.

II. System Model

We consider a system for eMBB-URLLC downlink scheduling. As shown in Figure 1, eMBB users and URLLC users are scheduled at different time scales. eMBB users are scheduled in slot, we consider N eMBB users, and eMBB users are indexed as $n \in \{1, 2, ..., N\}$. URLLC users are scheduled in mini-slot, we consider M URLLC packets, and packets indexed by $m \in \{1, 2, ..., M\}$. URLLC data packets have the characteristic of sporadic transmission. The arrival of URLLC packets is modeled as a Poisson distribution with mean rate λ. The generation probability of URLLC packets is modeled as a 0-1 distribution with probability P_{gen}. The downlink bandwidth is divided into L sub-bands indexed by $l \in \{1, 2, ..., L\}$, and a sub-band in the time of a mini-slot is a resource block (RB). Therefore, a mini-slot will have L RBs. The arriving URLLC packets have different delay requirements. The system needs to transmit the URLLC packets within the latency requirement.

A. Traffic model for eMBB users

The transmission rate of eMBB user n can be expressed by Equation (1), where $SNR_n(t) = \frac{P_n ||h_n(t)||^2 {d_n}^{\eta}}{N_0}$, B_n denotes the bandwidth resource allocated to eMBB user n. $h_n(t)$ denotes the fading channel power gain of eMBB user n at slot t, d_n denotes distance between eMBB user n and base station BS, η denotes the path loss exponent and N_0 is the noise power.

$$R\left(n\right) = B_n log_2 \left(1 + SNR_n\left(t\right)\right) \tag{1}$$

Because URLLC packets are scheduled on eMBB traffic, URLLC packets will puncture eMBB users, resulting in data rate loss of eMBB users. We consider that the scheduling of URLLC is punching scheduling, and model the data rate loss of eMBB users as a linear model, the data rate of the eMBB user after being punched is shown in Equation (2). where $num(n)$ represents the number of RBs allocated to eMBB user n, $num_m(n)$ represents the number of RBs punctured by URLLC packet m for eMBB user n.

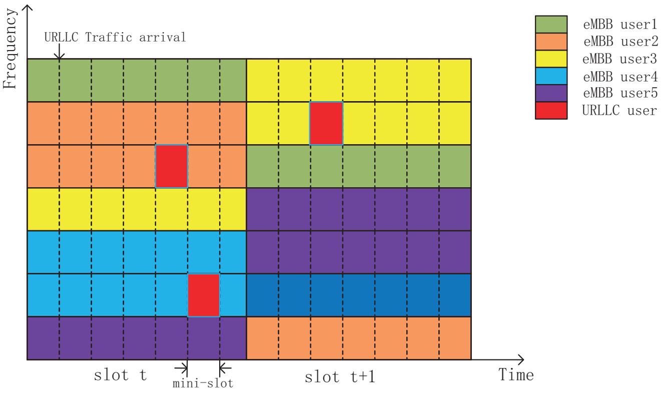

Fig. 1. Punching scheduling model in the joint eMBB-URLLC system

$$\phi\left(n\right) = \left(1 - \frac{\sum_m num_m\left(n\right)}{num\left(n\right)}\right) R\left(n\right) \tag{2}$$

B. Traffic model for URLLC packets

When considering the channel capacity, it is shown that the Shannon formula will no longer be applicable in the case of short packets [12], therefore, based on the finite block length coding theory [13], the capacity formula for URLLC is shown in Equation (3).

$$C_m = \frac{B_m}{ln2}\left[ln\left(1 + SNR_m\left(t\right) - \frac{\sqrt{V_m} Q_G^{-1}\left(\varepsilon_c\right)}{\sqrt{TTIB_m}}\right)\right] \tag{3}$$

We concentrate on the puncturing technique. In the punctured mini-slot, gNB allots zero power for eMBB user, and therefore, the interference cannot affect the URLLC packets. B_m represents the bandwidth occupied by the URLLC packet punch, TTI represents the time length of the mini-slot, and ε_c is the decoding error probability, which is directly related to the reliability of URLLC. $Q_G(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$ is the Gaussian Q function,$\sqrt{V_m} = \sqrt{1 - \frac{1}{\left(1 + \frac{g_m p_m}{N_0 B_w W_m}\right)^2}}$ is channel dispersion. For a URLLC packet, his transmission rate can be expressed as $\frac{pkt}{TTI}$, where pkt represents the data volume of the URLLC packet and TTI denotes the time of a mini-slot. We model the size of URLLC data packets as 32 bits, and one URLLC data packet can meet its transmission rate requirements by punching one RB.

III. Scheduling Processing of URLLC Packets

URLLC services are widely used in different business scenarios such as autonomous driving, power control, and industrial automation [14]. URLLC services have different latency and reliability requirements in these scenarios, as shown in Table I.

979-8-3503-0219-6/23/$31.00 ©2023 IEEE

TABLE I
LATENCY AND RELIABILITY REQUIREMENTS FOR COMMON URLLC SCENARIOS

Scenarios	Reliability/%	Air port delay/ms
Industrial Automation	99.9999	1
Power Distribution	99.9999	2-3
Remote Driving	99.999	3

TABLE II
KEY NOTATIONS

m/M	URLLC packets m/Total numbers of URLLC packets
n/N	eMBB user n/Total eMBB user numbers
l/L	Resource block l/ Total numbers of resource blocks
E_n/E	Resource allocation matrix for eMBB user n/Resource allocation matrix for total eMBB users
B_n/B_m	Bandwidth for eMBB user n/ Bandwidth of URLLC packets m
U_m/U	Punch matrix for URLLC packet m/Punch matrix of total URLLC packets
$num_m(n)$	Number of resource blocks punched by URLLC packet m for eMBB user n
$num(n)$	Number of resource blocks allocated by eMBB user n
λ	URLLC packet arrival mean rate
P_{gen}	Probability of URLLC packet generation
α_m	the maximum number of mini-slots that URLLC packet m can wait
$\phi(n)$	Data rate after eMBB user n is punched
Que	Queuing queue for URLLC packets
$R(n)$	Data rate after eMBB user n resource allocation

A. Latency Analysis of URLLC Packets

As can be seen from Figure 2, the transmission delay of the URLLC downlink in a 5G system consists of four main components respectively the queuing delay of the URLLC packets T_{que}, the base station processing delay T_{proc}^{gnb}, URLLC packets transmission delay T_{trans}, and the user's processing delay T_{proc}^{ue}. The transmission time delay of URLLC packets is Transmission Time Interval TTI.

According to the literature [15], the sum of the base station processing delay T_{proc}^{gnb} and user processing delay T_{proc}^{ue} is four times the transmission interval, denoted as T_{Proc}. We set the delay requirement of URLLC packets to τ. Then, the number of mini-slots for which URLLC packets can tolerate waiting is obtained from Equation (4):

$$\alpha_m = \frac{\tau}{TTI} - T_{Proc} - T_{trans} \tag{4}$$

For the three scenarios with latency requirements shown in Table I, we can calculate the number of mini-slots they can tolerate, which are denoted as α_{m1},α_{m2},α_{m3}.

B. Transmission Preprocessing of Arriving URLLC Packets

We create a queue. The length of the queue is the number of mini-slots that the URLLC packets with the largest delay can be queued. URLLC data packet in first position of the queue must be transmitted in the current mini-slot. URLLC data packet in second position of the queue must be transmitted in two mini-slots. And so on. If the URLLC packets have not been transmitted completely during the corresponding mini-slotS, the URLLC packets will violate its own delay requirements.

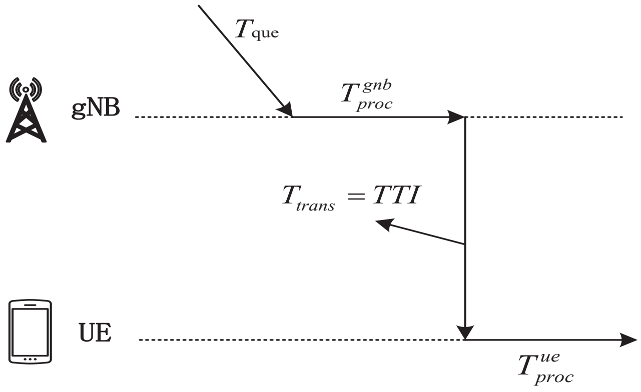

Fig. 2. Illustration of latency components for DL transmissions in NR

At the beginning of each mini-slot, we queue the URLLC packets that arrive. We calculate the average URLLC packet a transmitted at each position in the queue. We judge whether the number of URLLC data packets of the first position in the queue is greater than a, and if the number of URLLC packets of the first position in the queue is greater than a, we will transmit all URLLC packets of the first position in the queue. If it is less than a, the URLLC packets in the next position will also be transmitted in the current mini-slot until a URLLC packet is transmitted. The specific algorithm process is shown in Algorithm 1.

The proposed algorithm allocates the arriving URLLC packets with different delay requirements to multiple mini-slots for transmission. Because each mini-slot transmits a small amount of URLLC data packets, the RBs occupied by eMBB users with poor channel conditions can also meet the punching requirements of URLLC data packets, which can effectively reduce the loss of system throughput. The algorithm transmits URLLC within the delay requirement. It also effectively ensure the reliability of URLLC users under high loads.

Our proposed Algorithm 1 processes the queue in a single loop, so the complexity of Algorithm 1 is $O(len)$. len is the length of the queue.

IV. DYNAMIC SCHEDULING

We adopt an average allocation method for resource allocation among eMBB users in the system. All eMBB users in the system share the base station power and RBs equally. This allocation method can more conveniently observe the superiority of our URLLC scheduling algorithm.

In this section, we analyze the scheduling of URLLC packets in the current mini-slot. E is the resource allocation matrix of eMBB users. Its dimensions are N rows and L columns. E_n is the resource allocation matrix of eMBB user n. The matrix formula is as follows

$$E = \begin{bmatrix} E_1 \\ \vdots \\ E_n \\ \vdots \\ E_N \end{bmatrix} \tag{5}$$

$$E_n = \begin{bmatrix} 1 \cdots 1 \cdots 0 \cdots \end{bmatrix} \tag{6}$$

Algorithm 1 The Arriving URLLC Packets are Processed at Each Mini-Slot Using a Greedy Algorithm

```
Input: Arrival of URLLC packets $m_{total}$, Queue of URLLC packets $Que$
Output: Updated queue $Que$, Current mini-slot transmission packet $M$
The three delay-required URLLC packets that arrived in the current mini-slot are $m_1$,$m_2$,$m_3$, $m_1 + m_2 + m_3 = m_{total}$
Calculate the mini-slot that URLLC can be queued are $\alpha_{m1}$,$\alpha_{m2}$,$\alpha_{m3}$
Update $Que = [M_1, \cdots, M_{\alpha_{m1}} += m_1, \cdots, M_{\alpha_{m2}} += m_2, \cdots, M_{\alpha_{m3}} = m_3]$
Find the average number of URLLC packets to be transmitted per mini-slot $a = \frac{sum(Que)}{\alpha_{m3}}$
if $a$ is not an integer then
    round up
end if
if $a \leq M_1$ then
    transmit URLLC packets $M = M_1$
else
    $M = 0$
    for $i = 1, 2, 3, \cdots, \alpha_{m3}$ do
        if $M + M_i \leq a$ then
            transmit all $M_i$ URLLC packets in the current mini-slot and $M = M + M_i$
            update $M_i = 0$
        else
            Extract $a - M$ URLLC data packets from $M_i$ for current mini-slot transmission and $M = a$
            update $M_i = a - M$
        end if
    end for
end if
for each $i = 2, 3, \cdots, \alpha_{m3}$ do
    Update each element $M_{i-1} = M_i$
end for
update the last element $M_{\alpha_{m3}} = 0$
return $Que$,$M$
```

The elements $E_{n,l} \in \{0, 1\}$. When $E_{n,l} = 1$, it means that the l^th RB is assigned to eMBB user n, $E_{n,l} = 0$ means this RB is not assigned to eMBB user n. Because each RB can only be assigned to one eMBB user and all RBs will be assigned to eMBB users, so the constraint of the E matrix is:

$$E_{n,l} \in \{0, 1\}, \forall n \in \{1, ..., N\}, l \in \{1, ..., L\} \tag{7}$$

$$\sum_{n=1}^{N} E_{n,l} = 1, \forall n \in \{1, ..., N\}, l \in \{1, ..., L\} \tag{8}$$

U is the puncturing matrix of URLLC packets. Its dimensions are M rows and L columns. The elements $U_{m,l} \in \{0, 1\}$. When $U_{m,l} = 1$, it means that URLLC packet m punches l^{th} RB, and $U_{m,l} = 0$ means that the RB is not punched:

$$U = \begin{bmatrix} U_1 \\ \vdots \\ U_m \\ \vdots \\ U_M \end{bmatrix} \tag{9}$$

$$U_m = \begin{bmatrix} 1 \cdots 1 \cdots 0 \cdots \end{bmatrix} \tag{10}$$

A RB is punctured by at most 1 URLLC packet, and the U matrix satisfies the following constraints:

$$U_{m,l} \in \{0, 1\}, \forall m \in \{1, ..., M\}, l \in \{1, ..., L\} \tag{11}$$

$$\sum_{m=1}^{M} U_{m,l} \leq 1, \forall m \in \{1, ..., M\}, l \in \{1, ..., L\} \tag{12}$$

$\sum_m num_n(m)$ can be solved by the following equation, U_m^T denote transpose of U_m:

$$\sum_{m=1}^{M} num_n(m) = \sum_{m=1}^{M} E_n * U_m^T \tag{13}$$

The data rate after eMBB user n is punched is:

$$\phi(n) = (num_a(n) - \sum_{m=1}^{M} E_n * U_m^T) * R(n) \tag{14}$$

Therefore, the optimization objective is:

$$\max_{U} \sum_{n=1}^{N} \phi(n) \tag{15}$$

$$s.t. (11), (12), \tag{16}$$

The optimization problem is an integer programming problem. $U_{m,l}$ can only be 0 or 1. If we follow the conventional solution of the integer programming problem, we need to use RB as a variable, and the complexity of solving the problem will be relatively large. If the brute force solution is used, the complexity of URLLC scheduling is $\prod_{i=0}^{M-1}(L - i)$. Our optimization objective is to maximize the total data rate of eMBB users. The transmission rate of RB occupied by an eMBB user is same, so we only need to punch the minimum data rate of eMBB user to solve this optimization objective. Therefore, the complexity of Algorithm 2 is $O(M * N)$, M is the number of URLLC packets, N is the number of eMBB users. The specific algorithm flow is shown in Algorithm 2.

TABLE III
SIMULATION PARAMETERS

Parameters	Value
Number of eMBB users	5
Number of system resource blocks	120
RB Numbers Request Each packet	1
Transmission rate of eMBB users on one RB	$\{0.75, 0.875, 1, 1.25, 1.5\}$
Generation probability of URLLC traffic	$\{1/2, 1/5, 1/8\}$
Distribution of URLLC packets	Poisson distribution
Numerology	1

Algorithm 2 Punch Scheduling Algorithm

```
Input: URLLC Packet $m$
Output: Punch eMBB users $n$
while the number of URLLC packets $m$ is not 0 do
    $max = 0, Flage = 0$
    for all eMBB users do
        if eMBB user $n$ has remaining RBs and $\underset{U}{max} \sum_{n=1}^{N} \phi(n) > max$ then
            $max = \underset{U}{max} \sum_{n=1}^{N} \phi(n)$
            $Flage = n$
        end if
    end for
end while
return Puncture eMBB user $n$
```

V. Simulation Results

In the simulations, we consider a system with 120 RBs and 5 eMBB users. The eMBB users equally share the frequency band resources in the system, and each eMBB user is allocated 24 RBs. The transmission rate of each eMBB user on a RB is shown in Table III. We assume that the setting of eMBB remains unchanged at each slot.

As shown in Figure 3, we simulated 4000 slots to explore the RB distribution of URLLC packet punching. In this simulation, we set the probability of URLLC packet generation $P_{gen} = 1/2$, and the mean rate of the Poisson distribution of URLLC packets arrival $\lambda = 60$. The algorithm proposed in [5] immediately transmits URLLC data packets that arrive at the current mini-slot. In [5], it assumes that punching weights are in proportion to the percentage of resource (RP) that allocated to eMBB user n in each time slot. From Figure 3, it can be seen that the algorithm punches a large number of RBs with higher data rates. In [9], the authors employs queuing theory to divide all URLLC data packets into two mini-slots for transmission. Although this scheduling scheme can increase the proportion of RBs with lower data rates, which are punched by URLLC packets, the effect is not significant. Our algorithm provides a more comprehensive analysis of the number of mini-slots that URLLC packets can tolerate. Our algorithm effectively improves the proportion of RBs with smaller transmission rates, which are punched by URLLC data packets by dynamically distributing URLLC packets evenly across multiple mini-slots for transmission.

As shown in Figure 4, We simulate the relationship between the system throughput of eMBB users and the mean arrival rate of URLLC packets λ, when the generation probability of URLLC packets $P_{gen} = 1/2$. We can see that as the arrival rate of URLLC packets increases, the throughput of eMBB users in the system is gradually decreasing. Compared with the algorithms proposed in [5] and [9], our algorithm can effectively reduce the throughput loss of eMBB users in the system.

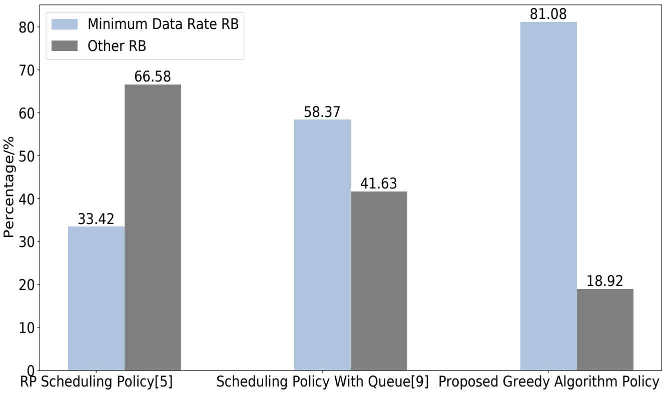

Fig. 3. Algorithm used VS. distribution of punched resource blocks

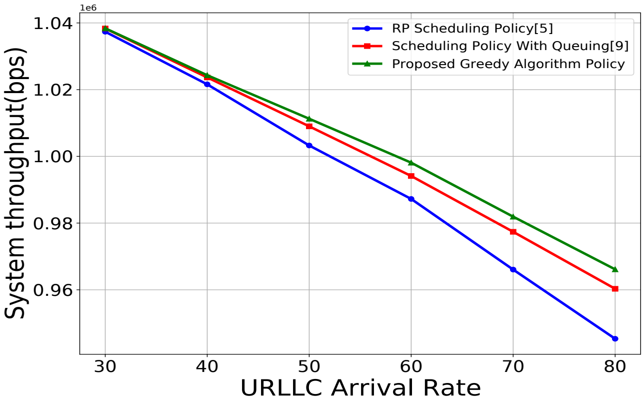

Fig. 4. eMBB throughput vs URLLC arrival rate λ

As shown in Figure 5, arrival rate of URLLC packets $\lambda = 70$. We simulate the relationship between the system throughput and the URLLC packet generation probability P_{gen}. It can be seen that the throughput of the system decreases as the probability of URLLC packet generation increases. From Figures 4 and 5, we can conclude that our algorithm can effectively reduce the throughput loss of the system when the arrival rate and generation probability of URLLC data packets increase.

As shown in Figure 6, we simulate the transmission of URLLC in the system when the load of URLLC is large. The probability of generating URLLC packets is $p_{gen} = 1/8$. From Figure 6, it can be seen that as the arrival rate of URLLC packets increases, the proposed algorithm can effectively ensure the reliability index of URLLC compared to the algorithm in [5]. Both our algorithm and the algorithm in [9]

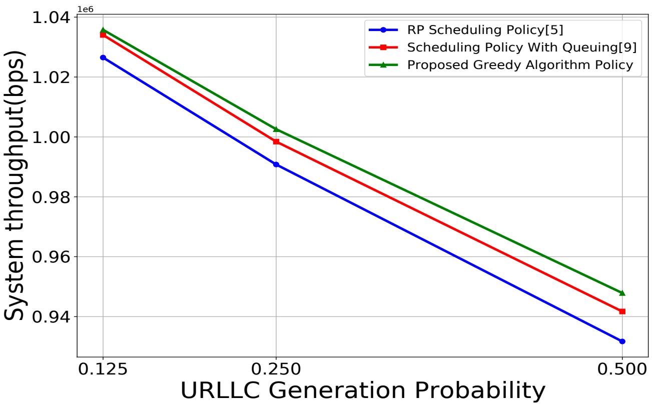

Fig. 5. eMBB throught vs URLLC generation probability P_{gen}

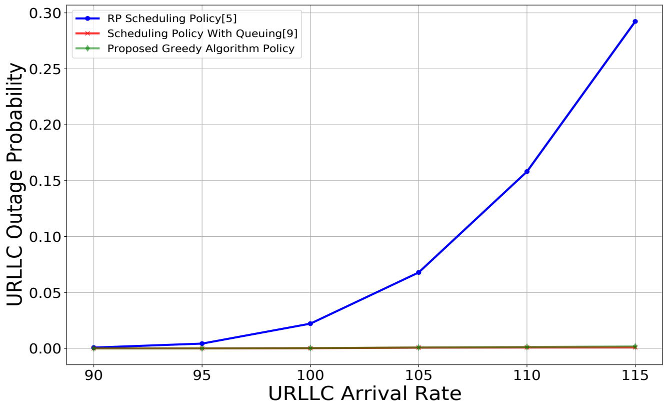

Fig. 6. URLLC outage probality vs URLLC arrival rate λ

can meet the reliability requirements of URLLC data packets under high loads. However, our algorithm's computational complexity is much smaller than the algorithm in [9].

VI. SUMMARY

In this paper, considering the different delay requirements of URLLC services in different scenarios, a greedy scheduling algorithm utilizing queuing theory is proposed. The algorithm dynamically distributes URLLC packets to multiple mini-slots for transmissions while guaranteeing the delay requirements of URLLC packets. The results show that compared [5] and [9], our algorithm can effectively reduce the throughput loss of eMBB users and ensure the reliability of URLLC under high loads.

REFERENCES

[1] E. Dahlman, S. Parkvall and J. Skold, 5G NR: The Next Generation Wireless Access Technology, New York, NY, USA:Academic, 2020.

[2] Y. B. Zikria, S. W. Kim, M. K. Afzal, H. Wang and M. H. Rehmani, "5G Mobile services and scenarios: Challenges and solutions", Sustainability, vol. 10, pp. 3626, 2018.

[3] H. Ji, S. Park, J. Yeo, Y. Kim, J. Lee and B. Shim, "Ultra-Reliable and Low-Latency Communications in 5G Downlink: Physical Layer Aspects," in IEEE Wireless Communications, vol. 25, no. 3, pp. 124-130, JUNE 2018.

[4] M. Bennis, M. Debbah and H. V. Poor, "Ultrareliable and Low-Latency Wireless Communication: Tail, Risk, and Scale," in Proceedings of the IEEE, vol. 106, no. 10, pp. 1834-1853, Oct. 2018.

[5] A. Anand, G. de Veciana and S. Shakkottai, "Joint Scheduling of URLLC and eMBB Traffic in 5G Wireless Networks," in IEEE/ACM Transactions on Networking, vol. 28, no. 2, pp. 477-490, April 2020.

[6] M. Alsenwi, N. H. Tran, M. Bennis, A. Kumar Bairagi and C. S. Hong, "eMBB-URLLC Resource Slicing: A Risk-Sensitive Approach," in IEEE Communications Letters, vol. 23, no. 4, pp. 740-743, April 2019.

[7] M. Almekhlafi, M. Chraiti, M. A. Arfaoui, C. Assi, A. Ghrayeb and A. Alloum, "A Downlink Puncturing Scheme for Simultaneous Transmission of URLLC and eMBB Traffic by Exploiting Data Similarity," in IEEE Transactions on Vehicular Technology, vol. 70, no. 12, pp. 13087-13100, Dec. 2021.

[8] A. K. Bairagi et al., "Coexistence Mechanism Between eMBB and uRLLC in 5G Wireless Networks," in IEEE Transactions on Communications, vol. 69, no. 3, pp. 1736-1749, March 2021.

[9] W. Zhang, M. Derakhshani and S. Lambotharan, "Stochastic Optimization of URLLC-eMBB Joint Scheduling With Queuing Mechanism," in IEEE Wireless Communications Letters, vol. 10, no. 4, pp. 844-848, April 2021.

[10] W. Zhang, M. Derakhshani, G. Zheng, C. S. Chen and S. Lambotharan, "Bayesian Optimization of Queuing-Based Multichannel URLLC Scheduling," in IEEE Transactions on Wireless Communications, vol. 22, no. 3, pp. 1763-1778, March 2023.

[11] B. Shi, F. -C. Zheng, C. She, J. Luo and A. G. Burr, "Risk-Resistant Resource Allocation for eMBB and URLLC Coexistence Under M/G/1 Queueing Model," in IEEE Transactions on Vehicular Technology, vol. 71, no. 6, pp. 6279-6290, June 2022.

[12] G. Durisi, T. Koch and P. Popovski, "Toward Massive, Ultrareliable, and Low-Latency Wireless Communication With Short Packets," in Proceedings of the IEEE, vol. 104, no. 9, pp. 1711-1726, Sept. 2016.

[13] S. Han et al., "Energy-Efficient Short Packet Communications for Uplink NOMA-Based Massive MTC Networks," in IEEE Transactions on Vehicular Technology, vol. 68, no. 12, pp. 12066-12078, Dec. 2019.

[14] X. Zhu, J. Wang, J. Li, H. Lu, X. Luo and Q. Lai, "An Approach to Transmitting URLLC Data with Different Latency Requirements over eMBB Services Based on Deep Reinforcement Learning," 2021 7th International Conference on Computer and Communications (ICCC), Chengdu, China, 2021, pp. 120-124.

[15] T. Fehrenbach, R. Datta, B. Göktepe, T. Wirth and C. Hellge, "URLLC Services in 5G Low Latency Enhancements for LTE," 2018 IEEE 88th Vehicular Technology Conference (VTC-Fall), Chicago, IL, USA, 2018, pp. 1-6.