Resource Allocation and Slicing Puncture in Cellular Networks with eMBB and URLLC Terminals Co-Existence

Yunzhi Zhao, Xuefen Chi, Lei Qian, Yuhong Zhu, and Fen Hou

***Abstract*—Ultra-reliable low-latency communication (URLLC) and enhanced mobile broadband (eMBB) are two types of services with delay quality of service (QoS) demands. Considering the random and sporadic URLLC packets arrival feature, slicing puncture is believed to be the suitable method to support co-existence communication scenario of eMBB and URLLC terminals. However, slicing puncture in a short time is not trivial, and needs accurate wireless resource scheduling. At the same time, it may damage the QoS of eMBB service. All of these make the QoS guaranteed scheduling problem more challenging. In this paper, we investigate the bandwidth, power allocations, slice puncturing problems to find the way satisfying services' individual QoS demands. For scheduling of eMBB service, we formulate a joint optimization problem for bandwidth and power allocations with long-term constraints of queues backlog. To solve this problem, we utilize the Lyapunov drift-plus-penalty method to establish the relationship between the long-term constraints and the short-term optimization problem, which means that the long-term constraints are gradually satisfied by the proposed strategy at each step. We further divide the short-term optimization problem into two subproblems, and adopt Block Coordinate Descent (BCD) algorithm to reduce the computation complexity. Then, we put forward a one-to-one matching method to solve the integer programming in resource block allocation and slicing puncture problems. Numerical results demonstrate that the proposed dynamic resource allocation and puncturing strategy (DRAPS) can solve the scheduling problem of eMBB and URLLC services in the presence of multiple randomnesses of channel state information (CSI) and URLLC packets arrival.**

***Index Terms*—Slicing puncture, resource allocation, URLLC and eMBB, Lyapunov optimization, block coordinate descent.**

I. Introduction

WITH the advent of the fifth generation and sixth generation (5G/6G) eras, many radically new applications emerge, such as smart healthcare, internet of vehicles, virtual reality and so on. These new applications can be broadly divided into three types of services: enhanced mobile broadband (eMBB), ultra-reliable and low latency communication (URLLC), and massive machine type communication (mMTC) [1-2]. To support these new applications, the 5G mobile networks are expected to provide 1000-fold network capacity [3]. Meanwhile, the consumptions of bandwidth and power are constantly increasing to satisfy the quality of service (QoS) demands. How to reduce resources consumption while guaranteeing the QoS of services is worth to be explored.

URLLC and eMBB terminals have different characteristics. Particularly, due to the enormous amount of traffic, eMBB terminals require wide bandwidth. In contrast, due to the ultra-reliable and low latency demands, URLLC terminals require that the system should schedule them immediately to fulfill the corresponding latency deadline. To meet the diverse requirements of different services, the network slicing method is proposed, which can provide customized service for different kinds of terminals by specialized configured slices [4]. Besides, the issue as how to properly schedule heterogeneous terminals has attracted much attention in recent years [5-8]. Since the QoS requirement of URLLC terminals is tight, the instantaneous required bandwidth is ultra-wide. In addition, the arrival packets of URLLC terminals are sporadic, if the system reserves bandwidth for URLLC terminals in advance, it will result in a waste of bandwidth resources. To solve this problem, a promising method named slicing puncture is proposed and these two heterogeneous services are scheduled on transmission time intervals (TTIs) of different durations, which means that the URLLC downlink arrival can be immediately scheduled in the next mini-slot and punctures the ongoing eMBB transmission [9-11]. The strategy of slice puncture will make adjustments of the previously configured slice resource and may spoil the QoS of eMBB terminals. The authors of [9] proposed three models to investigate the eMBB loss rate and formulated the eMBB and URLLC joint scheduling problem over two time scales. In [10], the authors explored the potential advantages of using non-orthogonal radio access network (RAN) resource for eMBB, mMTC and URLLC terminals and focused on developing a communication-theoretic model to capture the essential performance tradeoff for heterogeneous non-orthogonal multiple access (H-NOMA) and heterogeneous orthogonal multiple access (H-OMA). What's more, a risk-sensitive optimization problem was formulated to find the probability of each eMBB

This work is supported by the joint fund from the Ministry of Science and Technology of China and The Science and Technology Development Fund, Macau SAR (File No. 0066/2019/AMJ), by the Guangdong-Hong Kong-Macau Joint Laboratory Program (Project No. 2020B1212030009), by the Science and Technology Development Fund, Macau SAR (Project No. SKL-IOTSC(UM)-2021-2023), and by the Key Research and Development project of the Department of Science and Technology of Jilin Province (No. 20210201016GX).

Yunzhi Zhao and Fen Hou are with the State Key Laboratory of Internet of Things for Smart City, The Department of Electrical and Computer Engineering, University of Macau, Macao, China (email: yc17449@umac.mo and fenhou@um.edu.mo). Xuefen Chi, and Yuhong Zhu are with the Department of Communications Engineering, Jilin University, Changchun 130012, China (e-mail: chixf@jlu.edu.cn; yhzhu@jlu.edu.cn). Lei Qian is with the Tianjin Key Laboratory of Optoelectronic Detection Technology and System, School of Electrical and Electronic Engineering, Tiangong University, Tianjin 300160, China (e-mail: qianlei@tiangong.edu.cn) (Corresponding author: Xuefen Chi).

terminal being punctured [11] with the aid of Condition Value at Risk (CVaR) [12]. Moreover, the closed-form solutions for some optimization problems are hard to obtain due to the complexities of the New Radio (NR) physical layer. To avoid the obstacle and turn to search a feasible solution, a model-free deep reinforcement learning (DRL)-based solution was proposed, which aims to minimize the adverse impact of preemptive puncturing on eMBB terminals [13]. The above mentioned researches focused on the data loss rate of eMBB terminals caused by slicing puncture. In addition to data loss rate, delay is another significant performance metric. To our best knowledge, there is no research considering the delay QoS of eMBB explicitly under the premise of ensuring the reliability of URLLC terminals. Hence, the problem of delay guarantee under slicing puncture remains to be explored.

Jointly scheduling eMBB and URLLC terminals considering resources saving has many challenges. First of all, eMBB and URLLC terminals have distinct QoS requirements, and the delay requirement of URLLC terminals is more stringent than eMBB terminals. In this case, we consider the tight and statistical delay requirements for URLLC and eMBB terminals, respectively. However, statistical delay guarantee is difficult to realize without the prior knowledge of the probability distribution of system parameters. In addition, resource allocation has a great influence on the performance of QoS demand. Therefore, the tradeoff between the guarantee of QoS demand and resources saving is very important. Moreover, considering multiple stochastic characteristics such as random uncertain channel state information (CSI), random and sporadic arrival of URLLC packets, how to maintain the stability of our system in a long-term is important and challenging.

In this paper, we consider a downlink scenario where two types of heterogeneous services of eMBB and URLLC coexist. To save resources and satisfy different QoS requirements, we focus on bandwidth and power allocations, slicing puncture and delay guarantee in our system. Since the delay performance of eMBB may be damaged by the puncturing process, we propose an effective method for dynamic scheduling to reduce the influence. To the best of our knowledge, it is the first time to explore the joint scheduling problem of eMBB and URLLC terminals considering the tradeoff between QoS demands and resources saving in the cellular network. The main contributions of this paper are as follows.

- We jointly schedule eMBB and URLLC traffic through bandwidth, power allocations and slice puncture, which can satisfy the individual QoS requirements for URLLC and eMBB terminals, respectively. According to our proposed method, the system can allocate resources more flexible to tackle the resource competition problem caused by the scenario of eMBB and URLLC terminals co-existence.
- We transform the statistical delay requirement into an inequality about the expectation of queues backlog with the aid of Litter's law and Markov's inequality, which enables the realization of the statistical delay guarantee. Then an optimization problem with statistical delay constrains is formulated to minimize the utility function of eMBB terminals to tackle the demand of probabilistic constraint, we decompose the original problem into a series of deterministic short-term problems, which means that we bridge the statistical problem and the actual decision of each scheduling period with the help of Lyapunov optimization theory. Hence, we can meet the long-term constraints through the strategies of each scheduling period.
- To reduce the computational complexity, the transformed problem is further divided into two subproblems. We adopt an iteration method named Block Coordinate Descent (BCD) to solve the joint optimization problem. We propose a one-to-one matching approach to tackle the integer programming problems, which allocates the suitable resource blocks (RBs) to eMBB and selects sub-channels (SCs) belonging to the allocated RBs to URLLC terminals. The BCD method combined with one-to-one matching converges faster compared with genetic algorithm.
- We propose an online algorithm which refers to dynamic resource allocation combined with puncturing to solve the scheduling problem. In our proposed algorithm, we put forward a theoretical method to reveal the tradeoff between queues backlog and resources saving. Specifically, we introduce two control parameters, which are the values of weight factors in the Lyapunov function and the penalty function in our algorithm. More importantly, the control parameters can be adjusted to improve the performances that we concern. Particularly, we demonstrate the tendency of the performances varying with the control parameters and get upper bound of the expectation of queues backlog and utility function, respectively. The experimental results verify our conclusions.

The remainder of this paper is organized as follows. In section II, we introduce the system model about network structure and the queues backlog. We propose a stochastic optimization problem with long-term constraints and transform the primal problem into short-term optimization problems in section III. In addition, we elaborate the puncturing strategy for URLLC terminals in section IV and analyze the performance of the proposed algorithm in section V. In section VI, we present our simulation results. Finally, we make a conclusion in section VII. The notations mainly used in this paper are summarized in Table I.

II. System Model

A. Network Scenario

As shown in Fig. 1, we consider a network scenario with one base station (BS) and two types of terminals, i.e., eMBB terminals and URLLC terminals. We focus on the downlink transmission, where all terminals are scattered randomly around the BS. Let $\mathcal{E} = \{1, \ldots, E\}$ and $\mathcal{U} = \{1, \ldots, U\}$ denote the set of eMBB and URLLC terminals. Since URLLC arrival packets are sporadic [11], the set of URLLC terminals are divided into active URLLC terminals and inactive URLLC terminals, denoted by $\mathcal{U}^a$ and $\mathcal{U}^i$, respectively. The number of packets to be sent in the buffer of inactive URLLC terminals is zero.

B. Channel Model for eMBB Terminals

We assume that the BS perfectly knows the CSI of eMBB terminals from the pilot signal, and the channel power gain

TABLE I
MAIN SYMBOLS AND NOTATIONS

Symbols	Notations
$\mathcal{E}$	The set of eMBB terminals
$\mathcal{U}$	The set of URLLC terminals
$\mathcal{M}$	The set of RBs
$\mathcal{U}^a$	The set of active URLLC terminals
$\mathcal{U}^i$	The set of inactive URLLC terminals
e	The index of the eMBB terminal
u	The index of the URLLC terminal
m	The index of the RB
L_i	The distance between terminal i, $i \in \{e, u\}$
g_i	The channel fading of terminal i, $i \in \{e, u\}$
B_i	The bandwidth allocated to terminal i, $i \in \{e, u\}$
P_i	The transmit power of terminal i, $i \in \{e, u\}$
λ_e	The mean of the Poisson distribution for eMBB terminals
t	The index of the timeslot
t_{ms}	The index of the mini-slot, $t_{ms} \in \{1, ..., 7\}$
T_s	The duration of one timeslot
T_{ms}	The duration of one mini-slot
$I_{e,m}(t)$	The indicator of RB's allocation
$SC^p_{e,u}$	The number of SCs that the u^{th} URLLC terminal punctures e^{th} eMBB terminal
$Q_e(t)$	The queues backlog of the e^{th} eMBB terminal at timeslot t
$R_e(t)$	The departure rate of the e^{th} terminal at timeslot t
$A_e(t)$	The arrival rate of the e^{th} eMBB terminal at timeslot t
$A_u(t_{ms})$	The arrival rate of the u^{th} URLLC terminal at mini-slot t_{ms}
d_e	The delay requirement of the e^{th} eMBB terminal
δ_e	The threshold of the queue backlog of the e^{th} eMBB terminal
p_u	The packet sending probability of the u^{th} URLLC terminal
$I_{e,m}$	The indicator of RB allocation between the e^{th} eMBB terminal and the m^{th} RB
$\mathcal{F}_1$	The utility function in $\mathcal{Q}_1$ and $\mathcal{Q}_2$
$\mathcal{F}_2$	The utility function in $\mathcal{Q}_3$, $\mathcal{Q}_{3A}$, $\mathcal{Q}_{3B}$ and $\hat{\mathcal{Q}}_{3A}$
$G_e(t)$	The virtual queue of the e^{th} eMBB terminal at timeslot t
V_1, V_2	The control parameters in the DPP method
n	The index of iteration
ω_e	The number of RB allocated to the e^{th} eMBB terminal in RB matching problem of BCD algorithm
$D_e(t)$	The e^{th} eMBB terminal's degree of demand for another RB in RB matching problem
$SC^p_{e,u}(t_{ms})$	The number of the e^{th} eMBB terminal's SCs punctured by u^{th} URLLC terminal
$SC_u(t_{ms})$	The number of SCs punctured by the u^{th} URLLC terminal
$l^{ms}_e(t_{ms})$	The rate loss caused by puncture of the e^{th} eMBB terminal in mini-slot t_{ms}
$R^r_e(t_{ms})$	The maximum bits that can be transmitted from mini-slot t_{ms} to mini-slot 7
U_f	The utility function in $\mathcal{Q}_4$

can be expressed as

$$g_e = GL_e^{-a}|h_e|^2, \quad \forall e \in \mathcal{E}, \tag{1}$$

where G denotes the path loss constant, L_e denotes the distance between the BS and e^{th} eMBB terminal, and a is the path loss exponent. The small-scale fading h_e of eMBB terminals satisfies the distribution of $\mathcal{CN}(0,1)$. Hence, the data rate can be expressed as

$$r_e = B_e \log_2\left(1 + \frac{P_e g_e}{B_e \sigma^2}\right) \quad \text{bits/s}, \quad \forall e \in \mathcal{E}, \tag{2}$$

where P_e is the transmit power, B_e and σ^2 denote the allocated bandwidth to the e^{th} eMBB terminal and the Gaussian noise power spectral density, respectively.

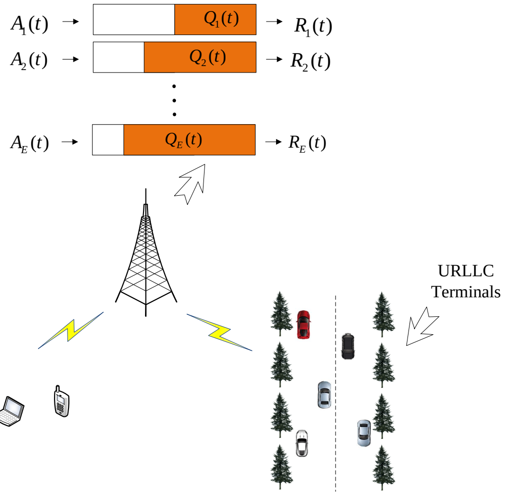


Fig. 1. The specific scenario in cellular networks.

C. Channel Model for URLLC Terminals

For URLLC terminals, we take vehicles as an example, which need high reliability and low delay service such as the transmission of information interaction between vehicles. Due to the fast movement of vehicles, we consider the imperfect estimation of CSI for URLLC terminals in this paper, and the small-scale fading for URLLC terminals is greatly influenced by the Doppler shift. The channel power gain of URLLC terminals can be expressed as [14]

$$g_u = GL_u^{-a}|h_u + \sqrt{1-\theta^2}\varsigma|^2, \quad \forall u \in \mathcal{U}, \tag{3}$$

where L_u denotes the distances between the BS and u^{th} URLLC terminal. The parameters h_u and ς are subject to $\mathcal{CN}(0,1)$. According to the Jakes statistical model [15], the coefficient θ $(0 < \theta < 1)$ describes the channel correlation.

In this paper, URLLC terminals require short data packets, and the transmission is not error-free. Hence, the Shannon capacity formula cannot be applied to characterize the achieved rate with a given error probability. We consider the short packet structure and transmission error rate. The achievable rate r_u with finite block-length l_u is expressed as [16]

$$\begin{aligned} r_u = & B_u \left\{ \log_2(1+\gamma_u) - \sqrt{\frac{\gamma_u(2+\gamma_u)}{(1+\gamma_u)^2 l_u ln(2)}} Q^{-1}(\varepsilon_u) \right\} \\ & \text{bits/s}, \quad \forall u \in \mathcal{U}, \end{aligned} \tag{4}$$

where $\gamma_u = \frac{P_u g_u}{B_u \sigma^2}$ denotes the signal-to-noise ratio (SNR) of the URLLC terminal, and ε_u is the finite block error rate (BLER). In addition, $Q^{-1}(.)$ denotes the inverse of the Gaussian Q function.

D. Dynamic Queue Model and Delay Requirement for eMBB Service

We consider a time-slotted model where t denotes the t^{th} timeslot and adopt the dynamic queue model to formulate the delay requirement of eMBB service. As shown in Figure 1, the BS maintains data queues for each eMBB terminal $e \in \mathcal{E}$, which is associated with the following attributes:

- Arrival rate $A_e(t)$: We assume the traffic arrival follows the Poisson distribution with the mean of λ_e bits/ms, i.e., $\mathbb{E}\{A_e(t)\} = \lambda_e$, and the arrival rate $A_e(t)$ is independent over each timeslot.
- Queue backlog $Q_e(t)$: The updated queue backlog $Q_e(t+1)$ of the e^{th} eMBB terminal can be obtained by queue backlog $Q_e(t)$, arrival data $A_e(t)$ and the departure rate $R_e(t)$ at the previous slot t, where $R_e(t) = r_e(t)T_s$ and T_s is the duration of one timeslot. Therefore, the queues backlog of each eMBB terminal can be updated as

$$Q_e(t+1) = \max\{Q_e(t) - R_e(t), 0\} + A_e(t), \ \forall e \in \mathcal{E}. \tag{5}$$

According to the Litter's law [24], the average delay is a function of queue length and arrival rate. We define the delay violation probability as the probability that the average queue delay is larger than the delay requirement d_e, and take a constraint about the delay violation probability. Then the delay requirement can be formulated as

$$\Pr\left\{\frac{Q_e(t)}{\lambda_e} \geq d_e\right\} \leq \epsilon, \quad \forall e \in \mathcal{E}, \tag{6}$$

where ϵ denotes the maximum delay violation probability. Further, we use Markov's inequality to rewrite (6) as

$$\Pr\left\{\frac{Q_e(t)}{\lambda_e} \geq d_e\right\} \leq \frac{\mathbb{E}\{Q_e(t)\}}{\lambda_e d_e}, \quad \forall e \in \mathcal{E}. \tag{7}$$

According to the large number law, when the time limits to infinite, the time average of the queue backlog is equal to the mathematical expectation. The delay requirement can be expressed as

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} Q_e(t) \leq \delta_e, \quad \forall e \in \mathcal{E}, \tag{8}$$

where $\delta_e = \lambda_e d_e \epsilon$ is the threshold of the queues backlog of the e^{th} eMBB terminal. So far, we have finished the transformation that using the queues backlog of eMBB terminals presents the delay requirement.

E. Scheduling for both eMBB and URLLC terminals

In this paper, we consider the coexistence of eMBB and URLLC services. Due to the random and sporadic arrival of URLLC packets, we first address the power and bandwidth allocations for eMBB scheduling without URLLC packets arrival in Section III. With the arrival of URLLC packets, we design efficient slicing puncture strategy to guarantee their low latency requirements in Section IV, where we jointly consider the scheduling for both eMBB and URLLC traffic such that their diverse QoS requirements can be satisfied and the impacts of URLLC traffic on eMBB service can be minimized.

III. JOINT POWER AND BANDWIDTH ALLOCATIONS OF EMBB SCHEDULING

In this section, we consider the power and bandwidth allocations simultaneously and formulate the problem as a stochastic optimization problem with long-term statistical constraints. Based on the Lyapunov optimization theory, we transform the long-term problem into a series of short-term deterministic optimization problems. However, the short-term problem is a kind of non-deterministic polynomial (NP) complete problem, which needs a lot of computing time to solve. Hence, we divide the short-term optimization problem into two subproblems, and propose an iterative algorithm to obtain a suboptimal solution.

A. Scheduling of eMBB Terminals

The bandwidth resource in a timeslot is divided into a set of RBs denoted by $\mathcal{M} = \{1, 2, ..., M\}$ and m denotes the index of RB. The bandwidth of each RB is denoted by B_{unit}. At each scheduling period, the system allocates some RBs to eMBB terminals to transmit information. Therefore, the binary assignment variable is denoted by $I_{e,m}(t)$, which is 1 if the m^{th} RB is allocated to the e^{th} eMBB terminal at the timeslot t, otherwise, it is 0. In addition, we assume that the CSI of eMBB terminals would not change rapidly and remain identical in each timeslot.

We aim at minimizing the bandwidth and power resources together. Hence, we define a utility function $\mathcal{F}_1$ which is a function of $B_e(t)$ and $P_e(t)$, i.e., $\mathcal{F}_1 = \sum_{e\in\mathcal{E}} B_e(t) + \alpha \sum_{e\in\mathcal{E}} P_e(t)$ and $B_e(t) = \sum_{m=1}^{M} I_{e,m}(t) B_{unit}$.

Thus, the joint power and bandwidth allocation problem can be formulated as the minimization problem $\mathcal{Q}_1$.

$$\mathcal{Q}_1 : \min_{I_{e,m}(t), P_e(t)} \mathcal{F}_1 \tag{9a}$$

$$\text{s.t.} \ \ I_{e,m}(t) = \{0, 1\}, \ \forall e \in \mathcal{E}, \forall m \in \mathcal{M}, \tag{9b}$$

$$\sum_{e\in\mathcal{E}} I_{e,m}(t) \leq 1, \ \forall m \in \mathcal{M}, \tag{9c}$$

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} Q_e(t) \leq \delta_e, \quad \forall e \in \mathcal{E}, \tag{9d}$$

$$0 \leq P_e(t) \leq P_{\max}, \ \forall e \in \mathcal{E}, \tag{9e}$$

$$R_e(t) \leq Q_e(t), \quad \forall e \in \mathcal{E}, \tag{9f}$$

where the constraints (9b) and (9c) guarantee that each RB is allocated to one eMBB terminal at most. The constraint (9d) guarantees the statistical delay of eMBB terminals. Remarkably, inequation (9d) represents the long-term constraints for queues backlog of eMBB terminals. If the queues backlog of constraints (9d) is stable, $\mathcal{Q}_1$ is feasible. In this paper, we assume that $\mathcal{Q}_1$ is always feasible, i.e., there exists at least one solution to satisfy all the constraints. Constraint (9e) represents that the transmit power is no more than the maximum transmit power $P_{\max}$. To save the bandwidth and power resources and maximize efficiency of resources utilization, we restrict the amount of departure rate which is denoted by constraint (9f).

However, problem $\mathcal{Q}_1$ cannot be directly solved through conventional optimization methods due to its long-term constraints. Fortunately, Lyapunov optimization theory provides us with a framework to solve this kind of problem. Therefore, we transform the original problem $\mathcal{Q}_1$ into a series of short-term optimization problems, which are described in the following subsections.

B. The Virtual Queues of eMBB Traffic

To tackle the average queues backlog constraint (9d), we introduce the concept of virtual queues [26], which can transform the long-term constraints into the stability problem of virtual queues. We define a virtual queue denoted by $G_e(t)$ for each queue backlog. Each virtual queue's update equation can be given by

$$G_e\left(t+1\right)=\max\left\{G_e\left(t\right)+Q_e\left(t+1\right)-\delta_e,0\right\},\quad \forall e\in\mathcal{E}. \tag{10}$$

$Lemma$ 1 : When the virtual queue $G_e\left(t\right)$ is rate stable, which means

$$\lim_{T\to\infty}\frac{G_e\left(T\right)}{T}=0,\quad \forall e\in\mathcal{E}. \tag{11}$$

The long-term constraint (9d) will be satisfied. That is, constraint (9d) is equivalent to equation (11).

Proof: See Appendix A. ■

Hence, we complete the transformation from each long-term constraint to the pure virtual queue stability problem. According to Lemma 1, the long-term constraints are always equivalent to the constraints of virtual queues stability. Hence, the original problem $\mathcal{Q}_1$ can be transformed as follows

$$\mathcal{Q}_2:\min_{I_{e,m}(t),P_e(t)}\quad \mathcal{F}_1 \tag{12a}$$

$$\text{s.t.}\quad I_{e,m}(t)=\{0,1\},\ \forall e\in\mathcal{E},\forall m\in\mathcal{M}, \tag{12b}$$

$$\sum_{e\in\mathcal{E}}I_{e,m}(t)\le 1,\ \forall m\in\mathcal{M}, \tag{12c}$$

$$\lim_{T\to\infty}\frac{G_e\left(T\right)}{T}=0,\quad \forall e\in\mathcal{E}, \tag{12d}$$

$$0\le P_e\left(t\right)\le P_{\max},\quad e\in\mathcal{E}, \tag{12e}$$

$$R_e(t)\le Q_e(t),\quad \forall e\in\mathcal{E}. \tag{12f}$$

C. Transformed Short-term Optimization Problems

Next, to solve the problem of virtual queue stability, i.e., $\mathcal{Q}_2$, we transform the problem $\mathcal{Q}_2$ into a series of short-term optimization problems by the drift-plus-penalty (DPP) method. We use notation Θ to represent the virtual queues vector of $G_e\left(t\right)$, i.e., $\Theta\left(t\right)\triangleq\left[G_1\left(t\right),\ G_2\left(t\right),\ldots G_E\left(t\right)\right]$. We assume that the vector $\Theta\left(t\right)$ evolves over $t\in\{1,\ 2,\ldots\}$. Then, the weighted-Lyapunov function is constructed as

$$L\left(\Theta\left(t\right)\right)\triangleq\frac{V_1}{2}\sum_{e\in\mathcal{E}}G_e\left(t\right)^2, \tag{13}$$

where $V_1(V_1>0)$ is a control parameter for the virtual queues. According to the Lyapunov function, the Lyapunov drift is defined as

$$\Delta\left(\Theta\left(t\right)\right)\triangleq\mathbb{E}\left\{L\left(\Theta\left(t+1\right)\right)-L\left(\Theta\left(t\right)\right)|\Theta\left(t\right)\right\}. \tag{14}$$

In addition, if $\Delta\left(\Theta\left(t\right)\right)$ has an upper bound, expression (11) can be proved. Utilizing the Lyapunov DPP method, we consider the objective function $\mathcal{F}_1$ in the optimization problem $\mathcal{Q}_2$ as a penalty function. Then, the DPP function is expressed as

$$\Delta\left(\Theta\left(t\right)\right)+V_2\mathbb{E}\left\{\mathcal{F}_1|\Theta\left(t\right)\right\}. \tag{15}$$

$Lemma$ 2 : The upper bound of the DPP expression (15) can be derived as

$$\begin{aligned}&\Delta\left(\Theta\left(t\right)\right)+V_2\mathbb{E}\left\{\mathcal{F}_1|\Theta\left(t\right)\right\}\\&\le\mathbb{E}\left\{\frac{V_1}{2}\sum_{e\in\mathcal{E}}\left[C_e+R_e\left(t\right)^2-2G_e\left(t\right)R_e\left(t\right)\right]|\Theta\left(t\right)\right\}\\&\quad+V_2\mathbb{E}\left\{\mathcal{F}_1|\Theta\left(t\right)\right\}.\end{aligned} \tag{16}$$

where $V_2(V_2>0)$ is the control parameter for the performance of the utility function.

Proof: See Appendix B. ■

Significantly, since all the previous states such as data queues of eMBB terminals and virtual queues are known at the timeslot t, $\sum_{e\in\mathcal{E}}C_e$ can be regarded as a fixed parameter for every timeslot. The upper bound of Lyapunov DPP only depends on $R_e(t)$, which is a function of $I_{e,m}(t)$ and $P_e(t)$. The parameters V_1 and V_2 can be adjusted to achieve the tradeoff between queues backlog and the utility function. For example, a large parameter V_1 means the queues backlog of eMBB terminals gets more attention.

According to the Lyapunov optimization framework, the upper bound (16) can be used as the objective function to develop the joint power and bandwidth allocation problem. From the given analysis, to satisfy the constraint (12d), we only need to minimize the upper bound of the DPP expression. In addition, by observing the state of the system at each timeslot, we transform the problem $\mathcal{Q}_2$ into the following optimization problem.

$$\mathcal{Q}_3:\min_{I_{e,m}(t),P_e(t)}\mathcal{F}_2 \tag{17a}$$

$$\text{s.t.}\quad I_{e,m}(t)=\{0,1\},\ \forall e\in\mathcal{E},\forall m\in\mathcal{M}, \tag{17b}$$

$$\sum_{e\in\mathcal{E}}I_{e,m}(t)\le 1,\ \forall m\in\mathcal{M}, \tag{17c}$$

$$0\le P_e\left(t\right)\le P_{\max},\quad e\in\mathcal{E}, \tag{17d}$$

$$R_e(t)\le Q_e(t),\quad \forall e\in\mathcal{E}, \tag{17e}$$

where $\mathcal{F}_2=\frac{V_1}{2}\sum_{e\in\mathcal{E}}\left[R_e\left(t\right)^2-2G_e\left(t\right)R_e\left(t\right)\right]+V_2\mathcal{F}_1$ .Obviously, due to the variables of $I_{e,m}(t)\in\{0,1\}$ and $P_e(t)$, the optimization problem $\mathcal{Q}_3$ is a mixed-integer nonlinear problem (MINLP), which is a non-convex optimization problem. Nowadays, heuristic algorithms such as Genetic Algorithm (GA) [17] are common methods to solve the MINLP problem. However, the computational complexity is very high to obtain a feasible solution. To reduce the complexity and tackle the MINLP problem $\mathcal{Q}_3$, we adopt the BCD Algorithm [18] and obtain a suboptimal solution through multiple iterations. Specifically, we divide the problem $\mathcal{Q}_3$ into two subproblems: (1) $\mathcal{Q}_{3A}$: Power allocation problem. (2) $\mathcal{Q}_{3B}$: Resource block matching problem.

1) Power allocation problem: For subproblem $\mathcal{Q}_{3A}$ at the n^{th} iteration, the variable of RB allocation $I_{e,m}(t),\forall e\in\mathcal{E},\forall m\in\mathcal{M}$ is given. Hence, we only need to optimize

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2022.3160647, IEEE Internet of Things Journal

the variable $P_e(t)$ and obtain the optimal solution at the n^{th} iteration, the subproblem of $\mathcal{Q}_{3A}$ at each iteration is expressed as

$$\mathcal{Q}_{3A} : \min_{P_e(t)} \quad \mathcal{F}_2 \tag{18a}$$
$$\text{s.t. } 0 \leq P_e\left(t\right) \leq P_{\max}, \quad e \in \mathcal{E}, \tag{18b}$$
$$R_e(t) \leq Q_e(t), \quad \forall e \in \mathcal{E}. \tag{18c}$$

Obviously, because the variable $I_{e,m}(t)$ is known, we make a variable substitution and regard the departure rate $R_e\left(t\right)$ with the logarithmic form as the variable of the objective function of $\mathcal{Q}_{3A}$. Hence, the first part of $\mathcal{F}_2$ is a quadratic function of $R_e(t)$ and the second part of $\mathcal{F}_2$ changes to $V_2 \sum_{e \in \mathcal{E}} B_e(t) + V_2 \alpha \sum_{e \in \mathcal{E}} (2^{\frac{R_e(t)}{T_s B_e(t)}} - 1) \frac{B_e(t)\sigma^2}{g_e} T_s$. Both of them are convex functions. In addition, the departure rate $R_e(t)$ is a monotone increasing function of $P_e(t)$ and the subproblem $\mathcal{Q}_{3A}$ can be rewrite as

$$\hat{\mathcal{Q}}_{3A} : \min_{R_e(t)} \quad \mathcal{F}_2 \tag{19a}$$
$$\text{s.t. } 0 \leq R_e(t) \leq X_e, \forall e \in \mathcal{E}, \tag{19b}$$

where $X_e = \max\{Q_e(t), B_e T_s \log_2(1 + \frac{P_{\max} g_e}{B_e \sigma^2})\}, \forall e \in \mathcal{E}$. Therefore, the problem $\hat{Q}_{3A}$ is a convex optimization problem which can be efficiently solved by standard convex optimization solvers like CVX [19]. After determining $R_e(t)$, the transmit power also can be calculated.

2) Resource block matching problem: After solving the problem $\mathcal{Q}_{3A}$ at the n^{th} iteration, we take the optimal power allocation value solved from the subproblem $\mathcal{Q}_{3A}$ denoted by $P_e^{n*}(t)$ into the subproblem $\mathcal{Q}_{3B}$ as known conditions. For the given allocated power solutions, the problem of RBs allocation can be optimized by solving the following problem.

$$\mathcal{Q}_{3B} : \min_{I_{e,m}(t)} \quad \mathcal{F}_2 \tag{20a}$$
$$\text{s.t. } I_{e,m}(t) = \{0, 1\}, \ \forall e \in \mathcal{E}, \forall m \in \mathcal{M}, \tag{20b}$$
$$\sum_{e \in \mathcal{E}} I_{e,m}(t) \leq 1, \ \forall m \in \mathcal{M}, \tag{20c}$$
$$R_e(t) \leq Q_e(t), \quad \forall e \in \mathcal{E}. \tag{20d}$$

There is no doubt that the resource block matching problem is a $0-1$ integer programming problem with the set of variable $I_{e,m}(t), \forall e \in \mathcal{E}, \forall m \in \mathcal{M}$.

Inspired by the matching theory [20], we change the problem $\mathcal{Q}_{3B}$ into a series of one-to-one matching problems. Specifically, we allocate each RB to a potential eMBB terminal step by step until one of these two cases happens: (a) All the RBs have been allocated. (b) The objective function value of $\mathcal{Q}_{3B}$ does not decline. All the eMBB terminals compete for one RB at the same time, and the system will select the eMBB terminal which can benefit the objective function $\mathcal{Q}_{3B}$ best. At the start of the one-to-one matching, the competing RB is the m^{th} RB and m is equal to 1. Then, the index of RB increases the number by one after selecting the optimal eMBB terminal for the last competing RB. In this case, we assume the e^{th} eMBB terminal is allocated ω_e RBs and define a metric denoted by $D_e(t)$ to measure the e^{th} eMBB terminal's degree of demand for another RB, which can be expressed as

$$D_e(t) = \frac{V_1}{2}[\hat{R}_e(t)^2 - \check{R}_e(t)^2 - 2G_e(t)(\hat{R}_e(t) - \check{R}_e(t))], \quad \forall e \in \mathcal{E}, \tag{21}$$

where $\hat{R}_e(t)$ denotes the departure rate and the number of RBs equals to the number of allocated RBs plus one. Hence, the number of RBs can be expressed as $\omega_e + 1$. In addition, $\check{R}_e(t)$ denotes the departure rate with allocated ω_e RBs. What's more, the smaller the value of $D_e(t)$, the more the eMBB terminal wants the RB. Therefore, we select the optimal eMBB terminal e^* according to the following expression

$$e^* = \underset{e}{\operatorname{argmin}}\, D_e(t), \forall e \in \mathcal{E}. \tag{22}$$

After selecting the optimal eMBB terminal e^* for the m^{th} RB, the number of allocated RB need to be updated and our system continues to solve the one-to-one matching problem for another RB based on (22). Hence, the $0-1$ integer programming problem $\mathcal{Q}_{3B}$ is simplified into a series of matching problem which is easy to be tackled. Fig. 2 simulates the process for the series of one-to-one matching problems. In the beginning, all of the eMBB terminals compete the 1^{th} RB and each ω_e equals to 0 at this moment. When the objective value of $\mathcal{Q}_{3B}$ achieves the minimum value, the BS terminates the algorithm of the proposed resource block matching problem and each eMBB terminal obtains own bandwidth resource.

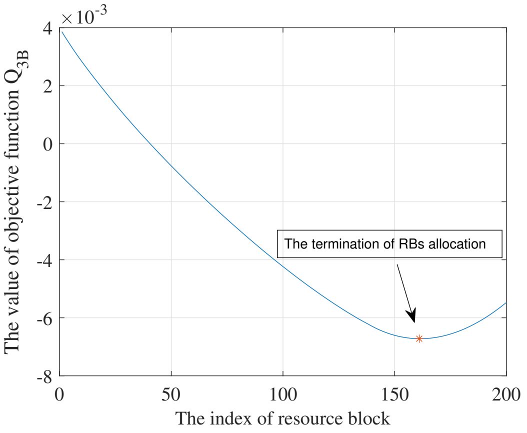


Fig. 2. The progress of solving the resource block matching problem.

According to the above explanation, our system allocates RBs and power resources to each eMBB terminal at the beginning of each timeslot t based on our proposed BCD algorithm which is summarized in Table II.

IV. Puncturing Strategy for URLLC Terminals

The random and sporadic URLLC packets will arrive at any time and our system has to schedule URLLC packets as soon as possible to guarantee the low latency requirement. In this section, we introduce the slicing puncture strategy and establish an optimization problem considering the queues backlog of eMBB terminals.

TABLE II
STRATEGY OF SOLVING PROBLEM $\mathcal{Q}_3$

```
Algorithm 1: BCD
1: Initialize bandwidth allocation $I$ and let $n = 0$, $\omega_e = 0$, $\forall e$;
2: while the objective value of $\mathcal{Q}_3$ does not converge:
3:     Power allocation problem:
           According to CVX solver, solve subproblem $\mathcal{Q}_{3A}^n$
           with given $I$ and obtain the optimal power allocation $P^*$;
4:     Resource block matching problem:
           Regard $P^*$ as constant in subproblem $\mathcal{Q}_{3B}$;
           while $m \le M$:
               Update $\hat{R}_e$ and $\check{R}_e$;
               Select the optimal eMBB terminal $e^*$ for the $m^{\text{th}}$ RB
               based on (22) and restricted by (20d);
               if the value objective function of $\mathcal{Q}_{3B}$ is non-decrease;
                   break;
               end if
               Update $m$, and $\omega_{e^*} = \omega_{e^*} + 1$;
           end while
5.     Update $I^*$ and $n$;
6.     Calculate the value of objective function $\mathcal{Q}_3$ based on $P^*$ and $I^*$;
7.end while
```

A. Slicing Puncture

According to NR Release-15 [21-22], a mini-slot in NR has a variable length including 2, 4 or 7 symbols. In our system, we consider the situation that each timeslot contains 7 mini-slots. We denote the index and duration of mini-slot as t_{ms} ($1 \le t_{ms} \le 7$) and T_{ms}, respectively. The packets belonging to URLLC terminals may arrive at any mini-slot duration and require a certain number of subcarriers (SCs) for transmission to satisfy its high reliability. To keep track of the pinpoint, special indication signals are generated and sent to the eMBB and URLLC terminals, which will be used for decoding [13]. Thus, the arriving URLLC packet is scheduled immediately to transmit in the next mini-slot on top of the ongoing eMBB transmissions, and the rate of the corresponding eMBB terminals will be affected by URLLC traffic in the mini-slot, which can be illustrated in Fig 3. Generally speaking, we control the queuing delay of URLLC terminals within a mini-slot duration.

To describe the situation of puncture in the mini-slot, we define a puncturing variable denoted by $SC_{e,u}^p(t_{ms})$ which represents the number of SCs that the u^{th} URLLC terminal punctures the e^{th} eMBB terminal in the mini-slot t_{ms}. To satisfy the delay requirement of URLLC terminals, each active URLLC terminal has to puncture eMBB terminals and obtains enough SCs for high reliable transmission. The number of SCs punctured by the u^{th} URLLC terminal can be calculated by

$$SC_u(t_{ms}) = \sum_{e \in \mathcal{E}} SC_{e,u}^p(t_{ms}), \quad \forall u \in \mathcal{U}^a. \tag{23}$$

According to the above section of eMBB scheduling, our system decides the allocation of bandwidth and power at each boundary of slot t based on the current situations. Hence, we only need to consider the puncturing strategy based on the given allocation strategy of eMBB terminals in each mini-slot to guarantee the URLLC terminals' demand of delay and reliability. Because the URLLC packet is characterized by small and sporadic data-package [25], the arrival packets of URLLC terminals are modeled as a random process. To simplify our system model, we consider a simple ON/OFF source model. Each URLLC terminal belongs to the active set with probability p_u and the arrival traffic of active URLLC terminals is denoted by A_u. Hence, the probability of active or inactive set can be expressed as

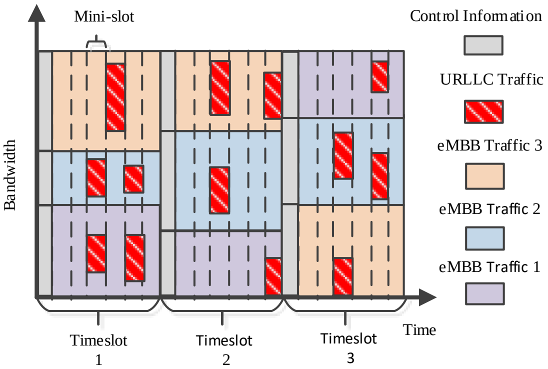


Fig. 3. The slicing puncture strategy for multiple eMBB terminals and URLLC terminals.

$$\Pr\{u \in \mathcal{U}^a\} = p_u, \quad \forall u \in \mathcal{U}, \tag{24}$$

$$\Pr\{u \in \mathcal{U}^i\} = 1 - p_u, \quad \forall u \in \mathcal{U}. \tag{25}$$

To satisfy the latency requirement of active URLLC terminals, if a URLLC packet arrives at the BS, our system will select appropriate radio resource, i.e., SC, belonging to eMBB terminals to transmit the URLLC packet in the next mini-slot. To guarantee the data transmission rate, our system has to satisfy the inequation constraint given by

$$r_u(t_{ms}) T_{ms} \ge A_u(t_{ms}), \quad \forall u \in \mathcal{U}^a, \tag{26}$$

which represents that the arrival packet will finish transmitting during the next mini-slot. Because the allocated bandwidth B_u for URLLC terminals in mini-slot t_{ms} depends on all eMBB terminals, $B_u(t_{ms})$ is a function of punctured bandwidth $B_{e,u}^p(t_{ms})$, where $B_{e,u}^p(t_{ms}) = SC_{e,u}^p(t_{ms}) \times 15000$ (The bandwidth of one SC is 15KHz). Therefore, we have the following expression.

$$B_u(t_{ms}) = \sum_{e \in \mathcal{E}} B_{e,u}^p(t_{ms}), \quad \forall u \in \mathcal{U}^a. \tag{27}$$

In addition, for inactive URLLC terminals, our system does not need to allocate SCs to them.

B. Optimization Problem for Puncturing Strategy

Due to the sporadic arrival of URLLC packets, eMBB terminals have no prior knowledge of the arrival of URLLC packets. Therefore, the puncturing strategy for URLLC terminals will impact the transmission rate of eMBB terminals. Thus, we aim to propose an optimal puncturing strategy which can not only satisfy the urgent requirement for URLLC service but also reduce the influence on eMBB service. We utilize the linear model [9] to measure the influence of puncture and model the rate loss of each eMBB terminal in mini-slot t_{ms} as

$$l_e^{ms}(t_{ms}) = \sum_{u\in\mathcal{U}^a} T_{ms} B_{e,u}^p(t_{ms}) \log_2(1+\gamma_e), \forall e \in \mathcal{E}, \quad (28)$$

where $\gamma_e = \frac{P_e g_e}{B_e \sigma^2}$. Then the whole rate loss of e^{th} eMBB terminal in timeslot t is given by

$$l_e(t) = \sum_{t_{ms}=1}^{7} l_e^{ms}(t_{ms}), \forall e \in \mathcal{E}. \quad (29)$$

When the URLLC packets arrive at mini-slot t_{ms}, the eMBB terminals' maximum bits that can be transmitted in the rest time of slot t is given by

$$R_e^r(t_{ms}) = (1 - \frac{t_{ms}-1}{7}) R_e(t) + \sum_{k=1}^{t_{ms}-1} l_e^{ms}(k), \forall e \in \mathcal{E}. \quad (30)$$

Considering the loose QoS requirement of eMBB terminals, we define a utility function denoted by U_f to measure the effect on QoS requirement of eMBB terminals based on $\mathcal{Q}_3$, which can be expressed as

$$\begin{aligned} U_f = & \sum_{e\in\mathcal{E}} \Bigg\{ (R_e^r(t_{ms}) - l_e^m(t_{ms}))^2 \\ & - 2G_e(t)(R_e^r(t_{ms}) - l_e^{ms}(t_{ms})) \Bigg\}. \end{aligned} \quad (31)$$

We determine the optimal strategy of puncture by solving the following optimization problem denoted by $\mathcal{Q}_4$.

$$\mathcal{Q}_4 : \min_{SC_{e,u}^p(t_{ms})} U_f \quad (32a)$$

$$\text{s.t. } r_u(t_{ms}) T_{ms} \geq A_u(t_{ms}), \quad \forall u \in \mathcal{U}^a, \quad (32b)$$

$$\sum_{u\in U^a} SC_{e,u}^p(t_{ms}) \leq SC_e(t_{ms}). \quad \forall e \in \mathcal{E}. \quad (32c)$$

Constraint (32c) denotes that the punctured SCs belonging to e^{th} eMBB terminal can not exceed the allocated SCs denoted by $SC_e(t_{ms})$ which can be obtained from the eMBB scheduling strategy. From the expression (31), $R_e^r(t_{ms})$ and $G_e(t)$ are constants at the beginning of mini-slot t_{ms}. The problem $\mathcal{Q}_4$ belongs to integer programming with the variable $SC_{e,u}^p(t_{ms})$. Similarly, the problem $\mathcal{Q}_4$ also can be regarded as a set of series of one-to-one matching problems, and each series of one-to-one matching problems determine the puncture for one URLLC terminal. Regardless of the constant in the expression (31), the metric which is used to select the optimal eMBB terminal can be expressed as

$$\begin{aligned} U_f^e(t_{ms}) = l_e^{ms}(t_{ms})^2 - 2R_e^r(t_{ms}) l_e^{ms}(t_{ms}) + 2G_e(t) l_e^{ms}(t_{ms}), \\ \forall e \in \mathcal{E}. \end{aligned} \quad (33)$$

We utilize $SC_{e,u}^{p*}(t_{ms})$ to represent the punctured SCs for u^{th} URLLC terminal. At each one-to-one matching problem for u^{th} URLLC terminal, we calculate the metric function $U_f^e(t_{ms})$ and fix the variable $SC_{e,u}^p(t_{ms})$ to $SC_{e,u}^{p*}(t_{ms})$ and $SC_{e,u}^{p*}(t_{ms}) + 1$, and select the optimal eMBB terminal based on

$$e^* = \operatorname*{argmin}_e \{\hat{U}_f^e(t_{ms}) - \check{U}_f^e(t_{ms})\}, \forall e \in \mathcal{E}, \quad (34)$$

where $\hat{U}_f^e(t_{ms})$ and $\check{U}_f^e(t_{ms})$ are calculated when $SC_{e,u}^p(t_{ms})$ equals to $SC_{e,u}^{p*}(t_{ms}) + 1$ and $SC_{e,u}^{p*}(t_{ms})$, respectively. Definitely, the selection of SC have to satisfy the constraint of (32c) and each series of one-to-one matching problems finishes when the constraint (32b) is satisfied. After finishing the puncture strategy and considering the influence on the rate of eMBB terminals, the expression (5) of the update expression of queue backlog can be rewritten as

$$\begin{aligned} Q_e(t+1) = \max\{Q_e(t) - R_e(t) + l_e(t), 0\} + A_e(t), \\ \forall e \in \mathcal{E}. \end{aligned} \quad (35)$$

It is worth noting that the rate loss owing to the puncturing strategy will be compensated at next time slot, which is benefited by the dynamic adjustment ability of Lyapunov optimization theory.

V. Performance Analysis

We tackle the eMBB scheduling problem based on Lyapunov optimization, and decompose the DPP problem denoted by $\mathcal{Q}_3$ into two subproblems. Furthermore, we propose an efficient method named one-to-one matching to solve the subproblem of bandwidth allocation and the problem of slicing puncture. In this section, we mainly analyze the performance of the proposed method.

A. Effectiveness and Stability Analysis of the Proposed eMBB Scheduling Strategy

We analyze the performance of the proposed eMBB scheduling strategy in this subsection. We assume that $\mathcal{Q}_1$ is feasible and the optimal solution is denoted by $\mathcal{F}_1^{opt}$. Because the objective function of $\mathcal{Q}_3$ is different from $\mathcal{Q}_1$, there exists a gap between $\mathcal{Q}_1$ and $\mathcal{Q}_3$, which will be discussed briefly in this subsection.

Compared to the objective function of the original problem $\mathcal{Q}_1$, the objective function of $\mathcal{Q}_3$ not only contains $\mathcal{F}_1$, but also contains the virtual queues of eMBB queues backlog.

Lemma 3 : The relationship between the problems $\mathcal{Q}_1$ and $\mathcal{Q}_3$ can be expressed as

$$0 \leq \mathbb{E}\{\mathcal{F}_1\} - \mathbb{E}\{\mathcal{F}_1^{opt}\} \leq \frac{B_{\text{cons}}}{V_2}, \quad (36)$$

where B_{cons} denotes a constant.

Proof: See Appendix C. ■

Inequality (36) suggests that the gap between $\mathbb{E}\{\mathcal{F}_1^{opt}\}$ and $\mathbb{E}\{\mathcal{F}_1\}$ decreases at the speed of $O(1/V_2)$. In other words, increasing the tradeoff parameter V_2 leads to $\mathcal{F}$ closer to the optimal solution $\mathcal{F}_1^{opt}$.

We assume that the virtual queues of eMBB queues backlog is rate stable. Because the Lyapunov function has an upper bound, the assumption is right. Based on the assumption and the fact $\mathbb{E}\left\{L\left(\Theta\left(0\right)\right)\right\}<\infty$, we derive the upper bound of expectation of queue backlog.

Lemma 4 : The upper bound for the exception of queue backlog can be expressed as follows

$$\mathbb{E}\left\{Q_e\left(T\right)\right\}\leq\sqrt{\frac{2(TB_{\text{cons}}+\mathbb{E}\left\{L\left(\Theta\left(0\right)\right)\right\})}{V_1}}+\delta_e,\quad\forall e\in\mathcal{E}. \tag{37}$$

Proof: See Appendix D. ■

Our proposed eMBB scheduling strategy with tradeoff parameters V_1 and V_2 provides an efficient method to decide the performance of $\mathcal{F}_1$ and the QoS requirements of eMBB terminals. Hence, the selection of these tradeoff parameters is decided by the demand of our system. For example, if the eMBB terminals in the system prefer a stricter delay, a larger V_1 is needed.

B. Convergence Analysis of the Proposed BCD Algorithm

As shown in problem $\mathcal{Q}_3$, there are two types of variables denoted by I and P about bandwidth and power allocations and all of the variables can be denoted by the set $\{I,P\}$. The proposed BCD algorithm contains a series of iterations, and each iteration includes two subproblems which are $\mathcal{Q}_{3A}$ and $\mathcal{Q}_{3B}$. For each subproblem, we keep the other type of variables fixed, and get the optimal solution through CVX solver or one-to-one matching method.

We define $\mathcal{Q}_{3A}^n(P^*)$ and $\mathcal{Q}_{3B}^n(I^*)$ as the optimal solution of subproblem $\mathcal{Q}_{3A}$ and $\mathcal{Q}_{3B}$ in n^{th} iteration, respectively. The solution P^* and I^* is optimal and better than any other corresponding policy. Therefore, we can get the following expressions

$$\mathcal{Q}_{3A}^n(P^*)\leq\mathcal{Q}_{3A}^n(P), \tag{38}$$

$$\mathcal{Q}_{3B}^n(I^*)\leq\mathcal{Q}_{3B}^n(I). \tag{39}$$

Moreover, the obtained optimal solution P^* is used as the input of the subproblem $\mathcal{Q}_{3B}$ at the same iteration and we also regard the optimal solution I^* as constant of subproblem $\mathcal{Q}_{3A}$ at the next iteration. We have the following expression which is given by

$$\begin{aligned}&\mathcal{Q}_3(\mathcal{Q}_{3A}^n,\mathcal{Q}_{3B}^n)\geq\mathcal{Q}_3(\mathcal{Q}_{3A}^{n*},\mathcal{Q}_{3B}^n)\geq\mathcal{Q}_3(\mathcal{Q}_{3A}^{n*},\mathcal{Q}_{3B}^{n*})\\&\geq\mathcal{Q}_3(\mathcal{Q}_{3A}^{(n+1)*},\mathcal{Q}_{3B}^{n*})\geq\mathcal{Q}_3(\mathcal{Q}_{3A}^{(n+1)*},\mathcal{Q}_{3B}^{(n+1)*}).\end{aligned} \tag{40}$$

The expression (40) indicates that the objective value of $\mathcal{Q}_3$ is non-increasing and the BCD algorithm is guaranteed to converge. Our proposed BCD algorithm needs to solve a convex optimization problem and a series of one-to-one matching problems in each iteration with low computational complexity. As shown in Fig. 4, we compare the proposed BCD algorithm with GA algorithm. Obviously, our proposed algorithm only needs several iterations to converge, which is more efficient than GA algorithm.

VI. Performance Evaluation

In this paper, we proposed the algorithm DRAPS (Dynamic Resource Allocation and Puncturing Strategy) to address the bandwidth and power resource allocations with the coexistence of eMBB and URLLC terminals. The flowchart of the proposed DRAPS algorithm is given in Fig. 5, which shows BCD algorithm for eMBB traffic, slicing puncture for URLLC traffic and their interaction. We have conducted extensive simulations to demonstrate the performance of the proposed algorithm, such as delay of eMBB, the packet sending probability of URLLC terminals, the impacts of the tradeoff parameters V_1, V_2 and son on. In addition, simulation results also validate our analysis and derivation.

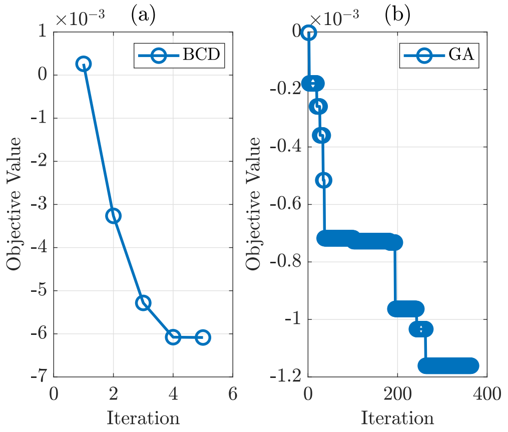


Fig. 4. Convergence of the proposed BCD and GA algorithm.

A. Simulation Settings

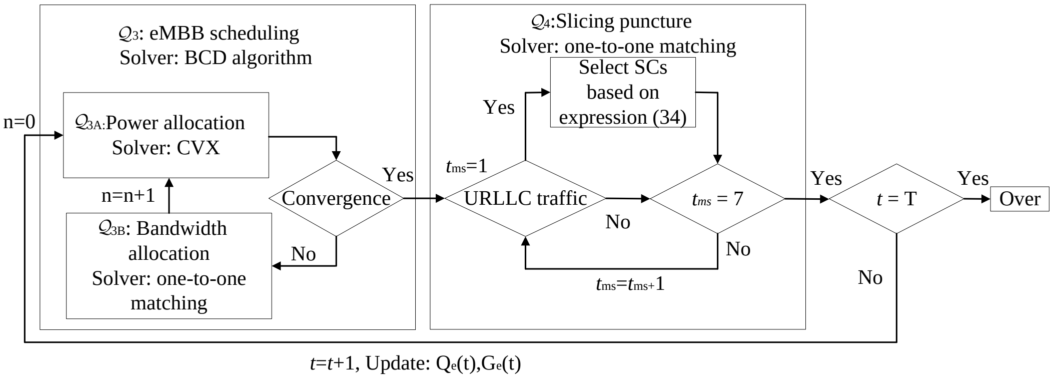
In our simulation, the BS is located at the center of the cell, and the radius of the cell is 500 meters. We suppose that terminals of eMBB and URLLC are randomly distributed around the BS. There are 8 eMBB terminals and 3 URLLC terminals in the cell. We assume that eMBB terminals' arrival rate A_e follows the Poisson distribution with a mean of λ_e ($\lambda_e=[80,40,20,80,60,30,50,80]$ kbit/T_s) during a long time T and the total number of RBs is 200. Furthermore, we assume that the maximum transmit power for terminals is $P_{\max}=0.5W$ and the path loss constant a is 3. The value of $A_u(t_{ms})$ is randomly selected from the set [32,200] bytes.

B. Effect of Delay Requirement and The Stability on The System

We consider the delay requirement of eMBB terminals, which will influence the performance of the proposed method. Meanwhile, we also verify the stability of the system according to the simulation results.

In Fig. 6, we analyze the effects of the sending probability of URLLC terminals and the delay requirement of eMBB terminals on the RB allocated to all eMBB terminals. Fig. 6 shows that the number of RBs increases gradually with the increase of the packet sending probability. Meanwhile, the delay requirement of eMBB terminals also has a great influence on the RB allocation. Different delay requirements of eMBB terminals result in different levels of long-term queues backlog requirements. The tighter the delay requirement of an eMBB terminal, the more resources the system will allocate

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2022.3160647, IEEE Internet of Things Journal



Q_3: eMBB scheduling
Solver: BCD algorithm
Q_{3A}:Power allocation
Solver: CVX
n=0
n=n+1
Q_{3B}: Bandwidth allocation
Solver: one-to-one matching
Convergence
Yes
No
Q_4:Slicing puncture
Solver: one-to-one matching
Select SCs based on expression (34)
Yes
t_{ms}=1
URLLC traffic
No
t_{ms} = 7
No
t_{ms}=t_{ms}+1
Yes
t = T
Yes
Over
No
t=t+1, Update: Q_e(t),G_e(t)

Fig. 5. The flowchart of the proposed DRAPS algorithm.

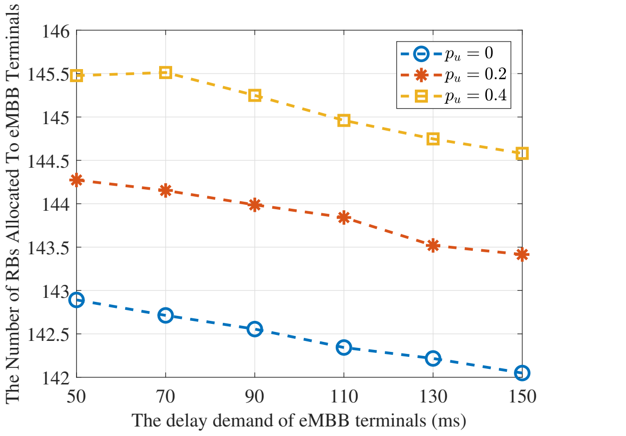


Fig. 6. The number of RBs allocated to eMBB terminals changes with the delay demand of eMBB terminals

to the terminal such that the individual QoS requirement can be satisfied.

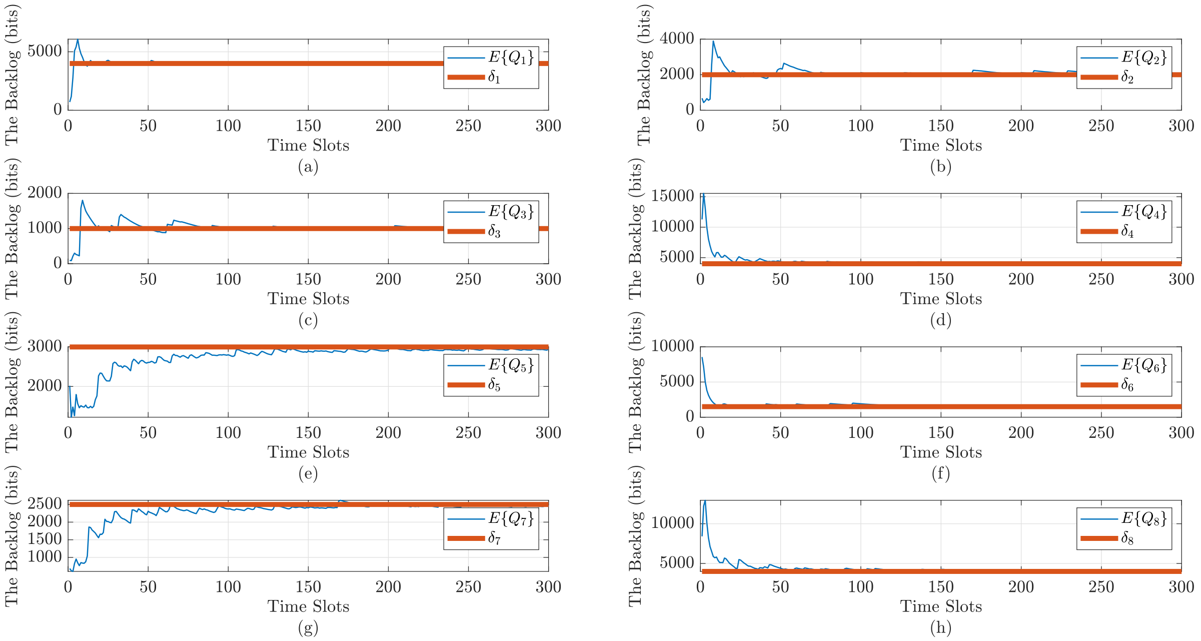
Fig. 7 shows the variation trend of the average queue backlog over time for eight eMBB terminals. The red line in each subgraph represents the threshold value of the queue backlog for the corresponding eMBB terminal. It is observed that the long-term average queue backlog denoted by $\lim\limits_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} Q_e(t)$ is no larger than the threshold δ_e, which means that the loose QoS requirement of each eMBB terminal is guaranteed through dynamic scheduling. More importantly, the departure rate $R_e(t)$ in each slot t is influenced by the random URLLC traffic. Because of the puncturing strategy, the virtual queue $G_e(t)$ and queue backlog $Q_e(t)$ are more than expected, our system will allocate more bandwidth and power resources to corresponding eMBB terminals at the next timeslot, and try its best to satisfy the constraint (9d). Therefore, by this dynamic scheduling method, our proposed DRAPS algorithm performs well in such uncertain situations.

C. Performance of The Proposed DRAPS Compared With Other Algorithms

When the packet of the URLLC terminal arrives, some appropriate SCs belonging to different eMBB terminals will be selected for the slicing puncture at the next mini-slot. Generally, if the packet sending probability of URLLC increases, the URLLC packets come more often and the eMBB terminals are more severely affected by slicing puncture. We compare the proposed DRAPS with other algorithms, the results are as follows.

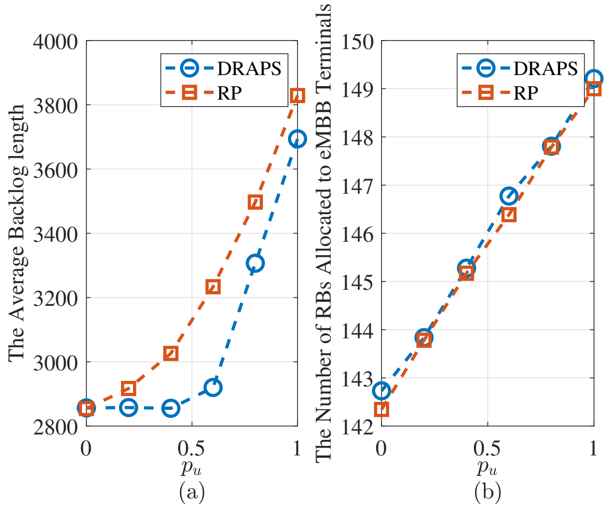
Fig. $8a$ shows the situation that the average queue backlog of eMBB terminals change with the packet sending probability of URLLC terminals. Meanwhile, we compare the proposed DRAPS algorithm with rate proportional (RP) algorithm [9] which is a broad class of mini-slot policy for puncture and the URLLC traffic is placed in mini-slots in proportion to the eMBB resource allocations. For the proposed DRAPS algorithm, the system selects SCs for slicing puncture considering the constraint (9d), which means that our system always considers QoS demand of eMBB terminals under situation of slicing puncture. Fig. $8b$ shows that our proposed slicing puncture strategy has the same number of allocated RBs compared with RP algorithm. However, in terms of the average queues backlog, our proposed DRAPS algorithm is superior to the RP algorithm regardless of different packet sending probabilities.

In Fig. 9, we investigate the effects of the packet sending probability p_u on the RBs allocated to all the eMBB terminals. As the parameter p_u increases, the number of allocated RBs also increases. We compare our eMBB scheduling strategy with Proportional Fair (PF) [28]. In our communication network, eMBB terminals experience different channel conditions. The PF scheduler not only aims to increase the throughput but also maintains the long-term allocation proportional fairness between eMBB terminals. From Fig. 9, compared to the PF algorithm, our proposed DRAPS algorithm requires less bandwidth resource to satisfy the heterogeneous service requirements.



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 7. The tendency of average queues backlog for the eMBB terminals in our system.



(a) (b)

Fig. 8. The allocated RBs and average queue backlog for our proposed DRAPS and RP algorithms.

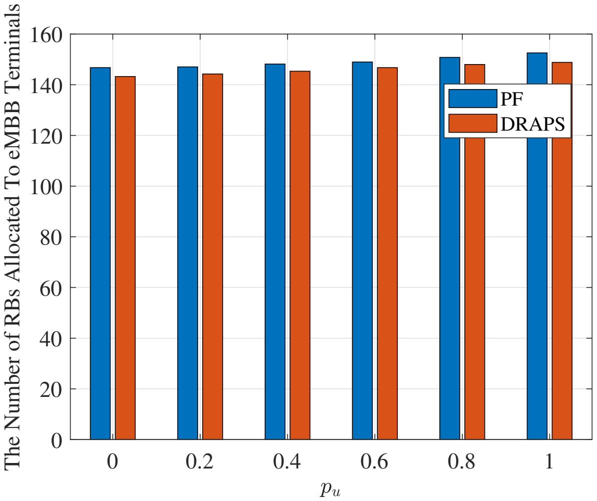


Fig. 9. The RBs allocate to eMBB terminals based on proposed DRAPS and PF algorithm, where $V_1 = 1 \times 10^{-13}$ and $V_2 = 1 \times 10^{-10}$.

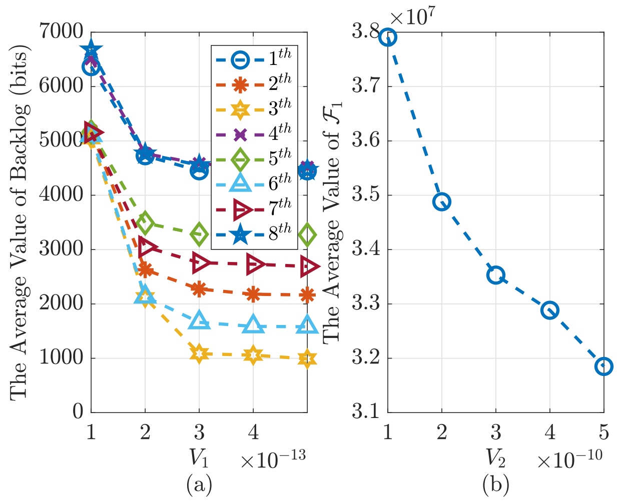
D. Effect of System Parameters

According to the above analysis in Section V, the system parameters such as V_1 and V_2 have vital roles in the system performance, which will describe the gap between problem $\mathcal{Q}_1$ and problem $\mathcal{Q}_3$. Hence, we prove our analysis via simulation.

We analyze the relationship between the average queues backlog of each eMBB terminal and the different control parameter V_1 in Fig. 10 (a). We can notice that as the parameter V_1 goes from 1×10^{-13} to 5×10^{-13}, the average queues backlog decrease. The reason is that our system pays more attention to the delay of eMBB terminals and makes more stringent limitations on the queues backlog with the increase of control parameter V_1. In Fig. 10 (b), we also investigate the variation tendency of our utility function $\mathcal{F}_1$ with different values of control parameter V_2. It is worth noting that the simulation results of the variation tendency verify our proofs expressed in expressions (36) and (37).

VII. Conclusion

In this paper, we consider the dynamic resources allocation with the coexistence of both eMBB and URLLC services. Due to the diverse QoS requirements of these two types of services and the sporadic arrival of URLLC packets, we propose BCD algorithm to efficiently allocate the power and



(a) (b)

Fig. 10. The system performance vary with the control parameters.

bandwidth resources for eMBB terminals to guarantee the statistical QoS requirement of eMBB terminals. With the arrival of URLLC traffic, we design an efficient resource allocation and slice puncture strategy DRAPS algorithm to guarantee the low latency of URLLC service and minimize the impacts on the performance of eMBB service as well. Simulation results demonstrate that, both the statistical QoS requirement of eMBB terminals and tight QoS requirement of URLLC terminals are satisfied with our proposed method. In addition, the computation complexity of our method is lower than the heuristic algorithm, e.g., GA.

APPENDIX A
PROOF OF LEMMA 1

For the virtual queue $G_e(t)$, we can rewrite its dynamic update expression as

$$\begin{aligned} & G_e(t+1) \\ & = \begin{cases} G_e(t) + Q_e(t+1) - \delta_e, \text{if } G_e(t) \geq \delta_e - Q_e(t+1) \\ \quad 0, \qquad\qquad\qquad\quad \text{if } G_e(t) < \delta_e - Q_e(t+1). \end{cases} \end{aligned} \tag{41}$$

Therefore, the difference between two adjacent timeslots is

$$\begin{aligned} & G_e(t+1) - G_e(t) \\ & = \begin{cases} Q_e(t+1) - \delta_e, & \text{if } G_e(t) \geq \delta_e - Q_e(t+1) \\ -G_e(t), & \text{if } G_e(t) < \delta_e - Q_e(t+1) \end{cases} \\ & = \max\{Q_e(t+1) - \delta_e, \ -G_e(t)\} \\ & \geq Q_e(t+1) - \delta_e. \end{aligned} \tag{42}$$

Summing both the left-hand side and the right-hand side of the inequality (42) from the timeslot 0 to the timeslot $T-1$ for $T \to \infty$. Further, we get

$$\begin{aligned} & \lim_{T\to\infty} G_e(T) - G_e(0) \geq \lim_{T\to\infty} \sum_{t=1}^{T} Q_e(t) - T\delta_e \\ & \geq \lim_{T\to\infty} \sum_{t=0}^{T-1} Q_e(t) - T\delta_e - Q_e(0) + \lim_{T\to\infty} Q_e(T). \end{aligned} \tag{43}$$

Dividing both sides of the inequality (43) by T, and using the fact that $G_e(0) < \infty$ and $Q_e(0) < \infty$. Hence, we have that $\lim\limits_{T\to\infty} \frac{G_e(0)}{T} = 0$ and $\lim\limits_{T\to\infty} \frac{Q_e(0)}{T} = 0$. We can get

$$\begin{aligned} & \lim_{T\to\infty} \frac{G_e(T)}{T} \\ & \geq \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} Q_e(t) - \delta_e + \lim_{T\to\infty} \frac{Q_e(T)}{T}. \end{aligned} \tag{44}$$

According to the rate stability theorem [27], the time averages of $A_e(t)$ and $R_e^{\max}(t)$ converge with probability 1 to finite constants $\widetilde{A}_e$ and $\widetilde{R}_e^{\max}$, i.e.,

$$\lim_{T\to\infty} \sum_{t=1}^{T} A_e(t) = \widetilde{A}_e, \quad \text{with probability } 1, \tag{45}$$

$$\lim_{T\to\infty} \sum_{t=1}^{T} R_e^{\max}(t) = \widetilde{R}_e^{\max}, \quad \text{with probability } 1. \tag{46}$$

If $Q_e(t)$ is rate stable, i.e., $\lim\limits_{T\to\infty} \frac{Q_e(T)}{T} = 0$, the system has to satisfy $\widetilde{A}_e \leq \widetilde{R}_e^{\max}$. According to the rate stability theorem, if $G_e(T)$ is rate stable, we can derive that

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} Q_e(t) \leq \delta_e. \tag{47}$$

We complete the proof of Lemma 1.

APPENDIX B
PROOF OF LEMMA 2

The drift of the Lyapunov function between two adjacent timeslots is

$$L(\Theta(t+1)) - L(\Theta(t)) = \frac{1}{2} \sum_{e\in\mathcal{E}} V_1 (G_e(t+1)^2 - G_e(t)^2). \tag{48}$$

To save resources, we always guarantee that the departure rate $R_e(t)$ can not exceed the queue length $Q_e(t)$. Therefore, the expression (5) can be rewrite as

$$Q_e(t+1) = Q_e(t) - R_e(t) + A_e(t), \quad \forall e \in \mathcal{E}. \tag{49}$$

Based on the virtual queue of queue backlog, we can derive that

$$\begin{aligned}
&G_e\left(t+1\right)^2 - G_e\left(t\right)^2 \\
&= \max\left\{G_e\left(t\right) + Q_e\left(t+1\right) - \delta_e, 0\right\}^2 - G_e(t)^2 \\
&\leq \left[G_e(t) + Q_e(t+1) - \delta_e\right]^2 - G_e(t)^2 \\
&= G_e\left(t\right)^2 + Q_e\left(t+1\right)^2 + {\delta_e}^2 + 2G_e\left(t\right)Q_e\left(t+1\right) \\
&\quad - 2G_e\left(t\right)\delta_e - 2Q_e\left(t+1\right)\delta_e - \ G_e\left(t\right)^2 \\
&= Q_e\left(t+1\right)^2 + {\delta_e}^2 + 2G_e\left(t\right)Q_e\left(t+1\right) \\
&\quad - 2G_e\left(t\right)\delta_e - 2Q_e\left(t+1\right)\delta_e \\
&= \left[Q_e\left(t\right) - R_e\left(t\right) + A_e\left(t\right)\right]^2 + {\delta_e}^2 \\
&\quad + 2G_e\left(t\right)Q_e\left(t+1\right) - 2G_e\left(t\right)\delta_e - 2Q_e\left(t+1\right)\delta_e \\
&\leq Q_e\left(t\right)^2 + A_e\left(t\right)^2 + R_e\left(t\right)^2 + 2Q_e\left(t\right)\left[A_e\left(t\right) - R_e\left(t\right)\right] \\
&\quad + {\delta_e}^2 + 2G_e\left(t\right)Q_e\left(t+1\right) - 2G_e\left(t\right)\delta_e - 2Q_e\left(t+1\right)\delta_e, \\
&\leq Q_e\left(t\right)^2 + A_e\left(t\right)^2 + R_e\left(t\right)^2 + 2Q_e\left(t\right)A_e\left(t\right) + {\delta_e}^2 \\
&\quad + 2G_e\left(t\right)\left[Q_e(t) - R_e(t) + A_e(t)\right].
\end{aligned} \tag{50}$$

With the given state of the last mini-slot, we have

$$\begin{aligned}
&E\left\{G_e\left(t+1\right)^2 - G_e\left(t\right)^2 \middle| \Theta\left(t\right)\right\} \\
&\leq C_e + R_e\left(t\right)^2 - 2G_e\left(t\right)R_e\left(t\right),
\end{aligned} \tag{51}$$

where $C_e = Q_e\left(t\right)^2 + A_e\left(t\right)^2 + 2Q_e\left(t\right)A_e\left(t\right) + {\delta_e}^2 + 2G_e\left(t\right)\left[Q_e\left(t\right) + A_e\left(t\right)\right]$. Taking conditional expectations to the Lyapunov function and substituting (51) into (48), we obtain the upper bound of the drift-plus-penalty expression as follows

$$\begin{aligned}
&\Delta\left(\Theta\left(t\right)\right) + V_2\mathbb{E}\left\{\mathcal{F}_1|\Theta\left(t\right)\right\} \\
&\leq \mathbb{E}\{\frac{V_1}{2}\sum_{e\in\mathcal{E}}\left[C_e + R_e\left(t\right)^2 - 2G_e\left(t\right)R_e\left(t\right)\right]\} \\
&\quad + V_2\mathbb{E}\left\{\mathcal{F}_1|\Theta\left(t\right)\right\}.
\end{aligned} \tag{52}$$

We complete the proof of Lemma 2.

Appendix C
Proof of Lemma 3

In practice, the arrival rate and the departure rate are constant for every timeslot. Hence, we assume $A_e\left(t\right) \leq A_{\max}$, $R_e\left(t\right) \leq R_{\max}$, $\delta_e \leq \delta_{\max}$, $\forall e$, $\forall t$. Because $Q_e(t)$ and $G_e(t)$ are rate stable, their maximum values are denoted by $Q_{\max}$ and $G_{\max}$. Therefore, we have

$$\begin{aligned}
C_e &= Q_e\left(t\right)^2 + A_e\left(t\right)^2 + 2Q_e\left(t\right)A_e\left(t\right) + {\delta_e}^2 \\
&\quad + 2G_e\left(t\right)\left[Q_e\left(t\right) + A_e\left(t\right)\right] \\
&\leq {Q_{\max}}^2 + {A_{\max}}^2 + 2Q_{\max}A_{\max} + {\delta_{\max}}^2 \\
&\quad + 2G_{\max}\left[Q_{\max} + A_{\max}\right] \\
&= C_1^{\max},
\end{aligned} \tag{53}$$

According to expression (53), and getting rid of the non-positive term on the right-hand side of the inequality (50), we rewrite expression (50) as

$$\begin{aligned}
&G_e\left(t+1\right)^2 - G_e\left(t\right)^2 \\
&\leq Q_e\left(t\right)^2 + A_e\left(t\right)^2 + R_e\left(t\right)^2 + 2Q_e\left(t\right)A_e\left(t\right) + {\delta_e}^2 \\
&\quad + 2G_e\left(t\right)\left[Q_e(t) - R_e(t) + A_e(t)\right]. \\
&\leq C_1^{\max} + {R_{\max}}^2.
\end{aligned} \tag{54}$$

If the problem $\mathcal{Q}_1$ is feasible, then for any $\psi > 0$ there always has a policy $\mathcal{F}_1$ that satisfies following expression

$$\mathbb{E}\{\mathcal{F}_1\} \leq \mathbb{E}\{\mathcal{F}_1^{opt}\} + \psi. \tag{55}$$

Hence,we can get the following expression

$$\begin{aligned}
&\mathbb{E}\{L\left(\Theta\left(t+1\right)\right) - L\left(\Theta\left(t\right)\right) + V_2\mathcal{F}_1\} \\
&\leq \mathbb{E}\{\frac{V_1}{2}\sum_{e\in\mathcal{E}}(G_e\left(t+1\right)^2 - G_e\left(t\right)^2)\} + V_2\mathbb{E}\left\{\mathcal{F}_1\right\} \\
&\leq \frac{V_1}{2}\left|E\right|\left(C_1^{\max} + {R_{\max}}^2\right) + V_2\mathbb{E}\left\{\mathcal{F}_1\right\} \\
&\leq B_{\text{cons}} + V_2\mathbb{E}\left\{\mathcal{F}_1\right\},
\end{aligned} \tag{56}$$

where $B_{\text{cons}} = \frac{V_1}{2}\left|E\right|\left(C_1^{\max} + {R_{\max}}^2\right)$ is a constant. Based on (56), we use telescoping sums over $t \in \{0, 1, \ldots, T-1\}$. Otherwise, we plug expression (55) into expression (56) and take a limit as $\psi \to 0$ and yield

$$\begin{aligned}
&\mathbb{E}\{L\left(\Theta\left(T\right)\right)\} - \mathbb{E}\{L\left(\Theta\left(0\right)\right)\} + TV_2\mathbb{E}\{\mathcal{F}_1\} \\
&\leq TB_{\text{cons}} + TV_2\mathbb{E}\left\{\mathcal{F}_1^{opt}\right\}.
\end{aligned} \tag{57}$$

Dividing expression (57) by $T \to \infty$, rearranging terms, and using the fact $E\{L\left(\Theta\left(0\right)\right)\} \leq \infty$, we yield

$$\begin{aligned}
&0 \leq \mathbb{E}\left\{\mathcal{F}_1\right\} - \mathbb{E}\left\{\mathcal{F}_1^{opt}\right\} \\
&\leq \lim_{T\to\infty}\frac{\mathbb{E}\{L\left(\Theta\left(0\right)\right)\} - \mathbb{E}\{L\left(\Theta\left(T\right)\right)\}}{TV_2} + \frac{B_{\text{cons}}}{V_2} \\
&\leq \frac{B_{\text{cons}}}{V_2}.
\end{aligned} \tag{58}$$

We complete the proof of Lemma 3.

Appendix D
Proof of Lemma 4

Similar as the proof of lemma 3, we prove that the average queue backlog of eMBB terminals varies with the control parameters, which means $\mathbb{E}\left\{Q_e\left(T\right)\right\}$ decreases with the increase of V_1. Meanwhile, we can obtain its upper bound.

Using the law of iterated expectations, we have

$$\mathbb{E}\{L\left(\Theta\left(T\right)\right)\} - \mathbb{E}\{L\left(\Theta\left(0\right)\right)\} \leq TB_{\text{cons}}. \tag{59}$$

Further, expression (59) is expressed as

$$\frac{V_1}{2}\mathbb{E}\left\{\sum_{e\in\mathcal{E}}G_e\left(T\right)^2\right\} - \mathbb{E}\left\{L\left(\Theta\left(0\right)\right)\right\} \leq TB_{\text{cons}}. \tag{60}$$

For the virtual queue of queue backlog, we have

$$\mathbb{E}\left\{\sum_{e\in\mathcal{E}}G_e\left(T\right)^2\right\} \leq \frac{2TB_{\text{cons}}}{V_1} + \frac{2\mathbb{E}\left\{L\left(\Theta\left(0\right)\right)\right\}}{V_1}. \tag{61}$$

Because $G_e\left(t\right)$ is a non-negative and independent for all the eMBB terminals. Based on the expression that

$$\mathbb{E}\left\{\sum_{e\in\mathcal{E}}G_e\left(T\right)^2\right\} \geq \mathbb{E}\left\{\left\{G_e\left(T\right)\right\}^2\right\} \geq \mathbb{E}\{G_e\left(T\right)\}^2, \tag{62}$$

we have

$$\begin{aligned}
\mathbb{E}\{G_e(T)\}^2 &\leq \frac{2TB_{\text{cons}}}{V_1} + \frac{2\mathbb{E}\left\{L\left(\Theta\left(0\right)\right)\right\}}{V_1} \\
&\leq \frac{2TB_{\text{cons}}}{V_1} + \frac{2\mathbb{E}\left\{L\left(\Theta\left(0\right)\right)\right\}}{V_1}.
\end{aligned} \tag{63}$$

For exprission $\mathbb{E}\{G_e(T)\}$, we have

$$\mathbb{E}\left\{G_e\left(T\right)\right\} \leq \sqrt{\frac{2(TB_{\text{cons}} + \mathbb{E}\left\{L\left(\Theta\left(0\right)\right)\right\})}{V_1}}. \qquad (64)$$

Because of $G_e\left(T\right) = \max\left\{G_e\left(T-1\right) + Q_e\left(T\right) - \delta_e, 0\right\}$, we get

$$G_e\left(T\right) \geq G_e\left(T-1\right) + Q_e\left(T\right) - \delta_e, \qquad (65)$$

and

$$Q_e\left(T\right) - \delta_e \leq G_e\left(T\right). \qquad (66)$$

Then, the upper bound of queue backlog can be expressed as

$$\mathbb{E}\left\{Q_e\left(T\right)\right\} \leq \sqrt{\frac{2(TB_{\text{cons}} + \mathbb{E}\left\{L\left(\Theta\left(0\right)\right)\right\})}{V_1}} + \delta_e. \qquad (67)$$

We complete the proof of Lemma 4.

REFERENCES

[1] X. Ge, S. Tu, G. Mao, C. -X. Wang and T. Han, ”5G Ultra-Dense Cellular Networks,” *IEEE Wireless Communications*, vol. 23, no. 1, pp. 72-79, February 2016.

[2] A. Osseiran et al., ”Scenarios for 5G mobile and wireless communications: the vision of the METIS project,” *IEEE Communications Magazine*, vol. 52, no. 5, pp. 26-35, May 2014.

[3] C. Wang et al., ”Cellular architecture and key technologies for 5G wireless communication networks,” *IEEE Communications Magazine*, vol. 52, no. 2, pp. 122-130, February 2014.

[4] X. Foukas, G. Patounas, A. Elmokashfi and M. K. Marina, ”Network Slicing in 5G: Survey and Challenges,” *IEEE Communications Magazine*, vol. 55, no. 5, pp. 94-100, May 2017.

[5] F. Song, J. Li, C. Ma, Y. Zhang, L. Shi and D. N. K. Jayakody, ”Dynamic Virtual Resource Allocation for 5G and Beyond Network Slicing,” *IEEE Open Journal of Vehicular Technology*, vol. 1, pp. 215-226, 2020.

[6] M. Yan, G. Feng, J. Zhou, Y. Sun and Y. Liang, ”Intelligent Resource Scheduling for 5G Radio Access Network Slicing,” *IEEE Transactions on Vehicular Technology*, vol. 68, no. 8, pp. 7691-7703, Aug. 2019.

[7] D. Wu, Z. Zhang, S. Wu, J. Yang and R. Wang, ”Biologically Inspired Resource Allocation for Network Slices in 5G-Enabled Internet of Things,” *IEEE Internet of Things Journal*, vol. 6, no. 6, pp. 9266-9279, Dec. 2019.

[8] V. N. Ha, T. T. Nguyen, L. B. Le and J. Frigon, ”Admission Control and Network Slicing for Multi-Numerology 5G Wireless Networks,” *IEEE Networking Letters*, vol. 2, no. 1, pp. 5-9, March 2020.

[9] A. Anand, G. De Veciana and S. Shakkottai, ”Joint Scheduling of URLLC and eMBB Traffic in 5G Wireless Networks,” *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, Honolulu, HI, 2018, pp. 1970-1978.

[10] P. Popovski, K. F. Trillingsgaard, O. Simeone and G. Durisi, ”5G Wireless Network Slicing for eMBB, URLLC, and mMTC: A Communication-Theoretic View,” *IEEE Access*, vol. 6, pp. 55765-55779, 2018.

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

[12] R. T. Rockafellar and S. Uryasev, ”Optimization of conditional value at risk,” in J. Risk, vol. 2, no. 3, pp. 21-41, 2000.

[13] Y. Huang, S. Li, C. Li, Y. T. Hou and W. Lou, ”A Deep-Reinforcement-Learning-Based Approach to Dynamic eMBB/URLLC Multiplexing in 5G NR,” *IEEE Internet of Things Journal*, vol. 7, no. 7, pp. 6439-6456, July 2020.

[14] Z. Wang, L. Liu, X. Wang and J. Zhang, ”Resource Allocation in OFDMA Networks With Imperfect Channel State Information,” *IEEE Communications Letters*, vol. 18, no. 9, pp. 1611-1614, Sept. 2014.

[15] Y. Polyanskiy, H. V. Poor and S. Verdu, ”Channel Coding Rate in the Finite Blocklength Regime,” *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2307-2359, May 2010.

[16] X. Li, L. Ma, Y. Xu and R. Shankaran, ”Resource Allocation for D2D-Based V2X Communication With Imperfect CSI,” *IEEE Internet of Things Journal*, vol. 7, no. 4, pp. 3545-3558, April 2020.

[17] K. Deb, A. Pratap, S. Agarwal and T. Meyarivan, ”A fast and elitist multiobjective genetic algorithm: NSGA-II,” *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 2, pp. 182-197, April 2002.

[18] M. Hong, M. Razaviyayn, Z. Luo and J. Pang, ”A Unified Algorithmic Framework for Block-Structured Optimization Involving Big Data: With applications in machine learning and signal processing,” *IEEE Signal Processing Magazine*, vol. 33, no. 1, pp. 57-77, Jan. 2016.

[19] S. Boyd and L. Vandenberghe, Convex Optimization. Cambridge, U.K.:Cambridge Univ. Press, 2004.

[20] D. Gale and L. S. Shapley, ”College admissions and the stability of marriage,” Amer. Math. Monthly, vol. 120, no. 5, pp. 386-391, 2013.

[21] J. Sachs, G. Wikstrom, T. Dudda, R. Baldemair and K. Kittichokechai, ”5G Radio Network Design for Ultra-Reliable Low-Latency Communication,” *IEEE Network*, vol. 32, no. 2, pp. 24-31, March-April 2018.

[22] S. Lien, S. Shieh, Y. Huang, B. Su, Y. Hsu and H. Wei, ”5G New Radio: Waveform, Frame Structure, Multiple Access, and Initial Access,” *IEEE Communications Magazine*, vol. 55, no. 6, pp. 64-71, June 2017.

[23] J. Navarro-Ortiz, P. Romero-Diaz, S. Sendra, P. Ameigeiras, J. J. Ramos-Munoz and J. M. Lopez-Soler, ”A Survey on 5G Usage Scenarios and Traffic Models,” *IEEE Communications Surveys & Tutorials*, vol. 22, no. 2, pp. 905-929, Secondquarter 2020.

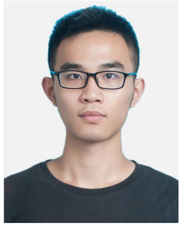
[24] J. D. C. Little and S. C. Graves, ”Little’s law,” in Building Intuition. New York, NY, USA: Springer, pp. 81-100, 2008.

[25] L. Feng, Y. Zi, W. Li, F. Zhou, P. Yu and M. Kadoch, ”Dynamic Resource Allocation With RAN Slicing and Scheduling for URLLC and eMBB Hybrid Services,” *IEEE Access*, vol. 8, pp. 34538-34551, 2020.

[26] Y. Li, M. Sheng, Y. Shi, X. Ma and W. Jiao, ”Energy Efficiency and Delay Tradeoff for Time-Varying and Interference-Free Wireless Networks,” *IEEE Transactions on Wireless Communications*, vol. 13, no. 11, pp. 5921-5931, Nov. 2014.

[27] M. J. Neely, ”Stochastic Network Optimization with Application to Communication and Queueing Systems”. San Mateo, CA, USA: Morgan & Claypool, 2010.

[28] H. Yin, L. Zhang and S. Roy, ”Multiplexing URLLC Traffic Within eMBB Services in 5G NR: Fair Scheduling,” *IEEE Transactions on Communications*, vol. 69, no. 2, pp. 1080-1093, Feb. 2021.

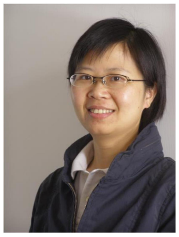


Yunzhi Zhao received the B.S. degree from the College of Communication Engineering, Jilin University, Changchun, China, in 2019. He is currently pursuing the Phd degree in the State Key Laboratory of IoT for Smart City, University of Macao. His research interests include resource allocation, Lyapunov optimization theory, ultra-reliable and low-latency communications, mobile edge computing and delay-quality of service (QoS) guarantees.

Xuefen Chi received the B.Eng. degree in applied physics from the Beijing University of Posts and Telecommunications, Beijing, China, in 1984, and the M.S. and Ph.D. degrees from the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, China, in 1990 and 2003, respectively. She was a Visiting Scholar with the Department of Computer Science, Loughborough University, U.K., in 2007, and the School of Electronics and Computer Science, University of Southampton, Southampton, U.K., in 2015. She is currently a Professor with the Department of Communications Engineering, Jilin University, China. Her research interests include machine-type communications, indoor visible light communications, random access algorithms, delay-QoS guarantees, and network modelling theory and its applications.

Lei Qian received the B.Eng., and Ph.D. degrees from the College of Communications Engineering, Jilin University, Changchun, China, in 2016 and 2021, respectively. Now she is with the Tianjin Key Laboratory of Optoelectronic Detection Technology and System, School of Electronic and Information Engineering, Tiangong University, as a lecturer. She was a visiting Ph.D. student with the School of Engineering, University of British Columbia, Canada, from 2019 to 2020, sponsored by the Chinese Scholarship Council. Her current research interests include delay QoS guarantee, effective capacity, visible light communications, resource allocation, physical layer security and channel estimation.

Yuhong Zhu received the B.S. degree in wireless communication from the Changchun Post and Telecommunications College, Changchun, China, in 1993, the M.S. degree in communication and information systems from the Beijing University of Posts and Telecommunications, Beijing, China, in 2000, and the Ph.D. degree in communication and information systems from Jilin University, Changchun, in 2012. From 2000 to 2005, he was a University Lecturer with the Department of Communications Engineering, Jilin University, where he is currently a Professor with the Department of Communications Engineering. His current research interests include wireless communication theory and applications, multimedia communications, and implementation and optimization of the algorithms in embedded systems.



Fen Hou received the Ph.D. degree in electrical and computer engineering from the University of Waterloo, Waterloo, ON, Canada, in 2008. She is currently an Associate Professor with the State Key Laboratory of IoT for Smart City and the Department of Electrical and Computer Engineering, and Guangdong-Hong Kong-Macau Joint Laboratory for Smart Cities, University of Macau, Macao, China. Her research interests include resource allocation and scheduling in broadband wireless networks, mechanism design and optimal user behavior in mobile crowdsensing networks, and mobile data offloading. Dr. Hou was a recipient of IEEE Globecom Best Paper Award in 2010 and the Distinguished Service Award in 2011. She served as the Co-Chair for INFOCOM 2014 Workshop on Green Cognitive Communications and Computing Networks, IEEE Globecom Workshop on Cloud Computing System, Networks, and Application 2013 and 2014, ICCC 2015 Selected Topics in Communications Symposium, and ICC 2016 Communication Software Services and Multimedia Application Symposium. She currently serves as an Associate Editor for IET Communications.