Admission Control and Network Slicing for Multi-Numerology 5G Wireless Networks

Vu Nguyen Ha, *Member, IEEE*, Ti Ti Nguyen, *Student Member, IEEE*, Long Bao Le, *Senior Member, IEEE*, and Jean-François Frigon, *Senior Member, IEEE*

***Abstract*—This letter studies the admission control and network slicing design for 5G New Radio (5G-NR) systems in which the total bandwidth is sliced to support the enhanced mobile broadband (eMBB) and ultra reliable and low latency communication (URLLC) services. We allow traffic from the eMBB bandwidth part to be overflowed to the URLLC bandwidth part in a controlled manner. We develop a mathematical framework to analyze the blocking probabilities of both eMBB and URLLC services based on which the network slicing and admission control is jointly optimized to minimize the blocking probability of the eMBB traffic subject to the blocking probability constraint for the URLLC traffic. An efficient iterative algorithm is proposed to deal with the underlying problem.**

***Index Terms*—Network slicing, 5G, new radio, numerology.**

I. Introduction

FUTURE mobile networks are expected to support a large number of wireless connections from different applications with diverse requirements including massive machine type communications (mMTC), eMBB, and URLLC. To this end, 5G-NR has proposed different types of physical resource blocks (PRBs) via the so-called flexible numerology [1]. As a result, each service can select a suitable numerology whose PRBs are assigned for its transmission to meet the requirements. Flexible numerology, thus, enables the 5G networks to effectively support heterogeneous services [2], [3]; however, it presents new challenges for resource management.

Development of access control mechanisms to effectively utilize the scarce bandwidth resource in 5G wireless systems is a major challenge which has been studied in several recent works [4]–[6]. While Popovski *et al.* propose network slicing strategies for three services mMTC–eMBB–URLLC in [4], the authors in [5] consider the scenario with mixed eMBB–URLLC traffic. While both papers consider the achievable transmission rates for eMBB and URLLC; only the work [4] imposes the throughput constraint for mMTC. The joint scheduling design for the eMBB and URLLC services is addressed in [6] where the URLLC traffic is scheduled on the eMBB bandwidth to meet the URLLC's low latency requirement and maximize the utility of eMBB traffic. However, the joint design of network slicing, numerology allocation, and admission control considering the 5G flexible numerology is not yet studied in these existing works.

This letter aims to fill this gap in the literature. In particular, we consider the admission control for the eMBB and URLLC services where the bounding control strategy [8] is employed to enable the eMBB blocked traffic to overflow to the URLLC's pre-assigned bandwidth. Then, an analytical framework is developed to determine the blocking probabilities (BPs) of the eMBB and URLLC traffic under this admission control strategy. Based on this analysis, we study the joint network slicing, numerology allocation, and admission control problem which aims to minimize the BP of the eMBB traffic subject to the BP constraint for the URLLC traffic. An efficient algorithm is proposed to solve this challenging problem. Finally, numerical studies are performed to validate the analytical model and demonstrate the efficiency of the proposed design. The simulation results also show that the joint design for network slicing, numerology allocation, and admission control can reduce the BP of eMBB traffic significantly while the BP constraint of URLLC traffic can be maintained. For easy reference, key notations used in this letter are summarized in Table I.

II. System Model

Consider a 5G new radio wireless network serving traffic flows generated from eMBB and URLLC services where a traffic flow represents a transmission request of the corresponding service with a data chunk to be transmitted over the wireless medium. We assume that the eMBB/URLLC traffic flows arrive according to different Poisson processes (PP) [7], [9], [10] with arrival rates λ_{e} and λ_{u}, respectively.[1] Additionally, the corresponding data lengths of the eMBB and URLLC flows follow general distributions with average values of $1/\mu_{\mathsf{e}}$ and $1/\mu_{\mathsf{u}}$, respectively.

We assume that the system bandwidth of W_{total} (MHz) is sliced into two bandwidth portions (BWPs) [11], denoted as L_{e} and L_{u}, which serve the eMBB and URLLC services, respectively. Let W_{e} and W_{u} (MHz) be the bandwidth of L_{e} and L_{u}, then $W_{\mathsf{e}} + W_{\mathsf{u}} = W_{\mathsf{total}}$. Various possibilities for numerology selection are allowed for the BWPs L_{e} and L_{u}. In particular, PRBs with high sub-carrier spacing

Manuscript received August 26, 2019; revised October 25, 2019; accepted December 11, 2019. Date of publication December 13, 2019; date of current version March 10, 2020. This work was supported in part by the National Sciences and Engineering Research Council of Canada under Grant RGPIN-2016-06401 and Grant RGPIN-2016-06550, and in part by Québec's Merit Scholarship Program for Foreign Students from Ministère de l'Éducation, de l'Enseignement Supérieur et de la Recherche du Québec, under Grant PBEEE-2018-262898. The associate editor coordinating the review of this letter and approving it for publication was J. Milizzo. *(Corresponding author: Vu Nguyen Ha.)*

V. N. Ha and J.-F. Frigon are with Poly-Grames Research Center, Polytechnique Montréal, Montréal, QC H3T1J4, Canada (e-mail: vu.ha-nguyen@polymtl.ca; j-f.frigon@polymtl.ca).

T. T. Nguyen and L. B. Le are with INRS-EMT, Université du Québec, Montréal, QC H5A1K6, Canada (e-mail: titi.nguyen@emt.inrs.ca; long.le@emt.inrs.ca).

Digital Object Identifier 10.1109/LNET.2019.2959733

[1]Typical use cases for URLLC and eMBB services are Internet of Things with small packets [9] and SPEED-5G virtual reality (VR) [10], respectively where the arrivals of their traffic flows are reported to follow the Poisson process.

TABLE I
LIST OF KEY NOTATIONS

Notations	Description
$\bar{\alpha}_{\mathsf{u}}$	Required minimum BP of the URLLC service
$\lambda_{\mathsf{e}}, \lambda_{\mathsf{u}}$	Arrival rates of eMBB and URLLC flows
$\lambda_{\mathsf{e}}^{\mathsf{u}}$	Arrival rates of eMBB overflow traffic
$1/\mu_{\mathsf{e}}, 1/\mu_{\mathsf{u}}$	Average data lengths of eMBB and URLLC flows
ρ_{e}, ρ_{u}, $\rho_{\mathsf{e}}^{\mathsf{u}}$	$\rho_{\mathsf{e}} = \lambda_{\mathsf{e}}\bar{t}_{\mathsf{e}}$, $\rho_{\mathsf{u}} = \bar{t}_{\mathsf{u}}\lambda_{\mathsf{u}}$, $\rho_{\mathsf{e}}^{\mathsf{u}} = \bar{t}_{\mathsf{e}}^{\mathsf{u}}\lambda_{\mathsf{e}}^{\mathsf{u}}$
ω, γ	IPP approximate parameters
Ω	$\Omega = \{W_{\mathsf{e}}, W_{\mathsf{u}}, g_{\mathsf{e}}, g_{\mathsf{u}}, K_{\mathsf{u}}\}$
Ω_1	$\Omega_1 = [W_{\mathsf{e}}, W_{\mathsf{u}}]$
Ω_2	$\Omega_2 = [g_{\mathsf{e}}, g_{\mathsf{u}}]$
g_{e}, g_{u}	Potential numerologies for L_{e}, L_{u}
K_{u}	Threshold of eMBB overflow traffic to BWP L_{u}
L_{e}, L_{u}	BWPs of eMBB and URLLC
W_{total}	Total system bandwidth
N_{e}, N_{u}	Numbers of SCs in BWPs L_{e} and L_{u}
r_{Hz}	Normalized data transmission rate per Hz
$R(g)$	$R(g) = r_{\mathsf{Hz}}W_0 2^g$
$\bar{t}_{\mathsf{e}}$, $\bar{t}_{\mathsf{u}}$	Average service time of eMBB and URLLC flows
$\bar{t}_{\mathsf{e}}^{\mathsf{u}}$	Average service time of eMBB overflow traffic in L_{u}
W_{e}, W_{u}	Bandwidth of L_{e} and L_{u}
W_0	Bandwidth of a PRB of 4G LTE system (0.18 MHz)

and short time slot duration are arranged for the BWP L_{u} to support the ultra-reliability and very low latency demands of URLLC applications while traffic flows from the eMBB service can adopt a numerology with the sub-carrier spacing larger than what are available in the 4G system but smaller than that required by the URLLC service. Therefore, we consider numerology selection options for these two services as $g_{\mathsf{e}} \in \{1, 2, 3, 4\}$ and $g_{\mathsf{u}} \in \{5, 6\}$ [1] where g_{e} and g_{u} are the potential numerologies for L_{e}, L_{u}, respectively.

Each BWP is divided into a number of sub-channels (SCs) each of which spans over a group of 12 contiguous sub-carriers in the frequency dimension. In 5G-NR, bandwidth of a SC corresponding to numerology g ($g = g_{\mathsf{e}}$ or g_{u}) is 2^g times of 180 kHz which is the bandwidth of a SC in the 4G LTE wireless system. Without loss of generality, we assume that one SC is assigned to serve one active flow[2] and the SC will be available for other flows when data transmission of the underlying flow is finished. Let $\lfloor . \rfloor$ stand for the floor operator, the numbers of SCs in BWPs L_{e} and L_{u} can be expressed, respectively as $N_{\mathsf{e}} = \lfloor W_{\mathsf{e}}/(W_0 2^{g_{\mathsf{e}}}) \rfloor$ and $N_{\mathsf{u}} = \lfloor W_{\mathsf{u}}/W_0 2^{g_{\mathsf{u}}} \rfloor$ where $W_0 = 0.18$ (MHz). We assume that URLLC data traffic is only transmitted over BWP L_{u} due to the strict URLLC requirements while eMBB data traffic can be transmitted over BWP L_{e} and L_{u} based on an admission control scheme as follows. When a new eMBB flow arrives, one available SC of L_{e} is picked to serve this transmission. If there is no available SC in BWP L_{e}, the flow is overflowed to and served by a SC in BWP L_{u} if there exists an available SC in BWP L_{u}. In addition, the maximum number of eMBB flows overflowed to BWP L_{u} is bounded by a pre-determined threshold K_{u}, i.e., $K_{\mathsf{u}} \leq N_{\mathsf{u}}$, to protect the quality of service (QoS) of the URLLC service [8].

III. BLOCKING PROBABILITY ANALYSIS

Let r_{Hz} (bits/s/Hz) be the normalized data transmission rate per Hz. Then, the average service time of one eMBB/URLLC flow in L_{e} and L_{u} can be expressed as $\bar{t}_{\mathsf{e}} = 1/(\mu_{\mathsf{e}} R(g_{\mathsf{e}}))$ and $\bar{t}_{\mathsf{u}} = 1/(\mu_{\mathsf{u}} R(g_{\mathsf{u}}))$, respectively where $R(g) = r_{\mathsf{Hz}} W_0 2^g$. Note that the service time of an overflowed eMBB flow admitted to BWP L_{u} decreases because of the larger bandwidth of a SC in BWP L_{u}. Then, its average service time can be expressed as $\bar{t}_{\mathsf{e}}^{\mathsf{u}} = \bar{t}_{\mathsf{e}} 2^{(g_{\mathsf{e}} - g_{\mathsf{u}})}$.

Let us define $\rho_{\mathsf{e}} = \lambda_{\mathsf{e}}\bar{t}_{\mathsf{e}} = \lambda_{\mathsf{e}}/(R(g_{\mathsf{e}})\mu_{\mathsf{e}})$. Then, the eMBB traffic in BWP L_{e} can be modeled as an *M/G/n/n* queue with the offered traffic intensity ρ_{e} and N_{e} servers. Hence, the BP of eMBB data flows in BWP L_{e} can be calculated from the Erlang B formula as follows [12]:

$$\mathsf{Pr}_{\mathsf{e,blk}}^{\mathsf{E}} = B(\rho_{\mathsf{e}}, N_{\mathsf{e}}) = \left(\frac{\rho_{\mathsf{e}}^{N_{\mathsf{e}}}}{N_{\mathsf{e}}!}\right) \Big/ \sum_{i=0}^{N_{\mathsf{e}}} \frac{\rho_{\mathsf{e}}^{i}}{i!}. \tag{1}$$

1) Overflow Traffic Approximation: Note that the overflow traffic does not follow PP [13]. To analyze the BP, the overflow flows from BWP L_{e} can be represented by an interrupted Poison process (IPP) with the arrival rate $\lambda_{\mathsf{e}}^{\mathsf{u}}$, the mean ON-time and OFF-time of the random switch being $1/\gamma$ and $1/\omega$, respectively as illustrated in Fig. 1 [14]. In addition, Kuczura in [14] has shown that an accurate approximation can be achieved if $\lambda_{\mathsf{e}}^{\mathsf{u}}$, γ, and ω are determined as follows:

$$\lambda_{\mathsf{e}}^{\mathsf{u}} = \rho_{\mathsf{e}} \frac{\delta_2(\delta_1 - \delta_0) - \delta_0(\delta_2 - \delta_1)}{(\delta_1 - \delta_0) - (\delta_2 - \delta_1)},$$
$$\omega = \frac{\delta_0}{\lambda_{\mathsf{e}}^{\mathsf{u}}}\left(\frac{\lambda_{\mathsf{e}}^{\mathsf{u}} - \rho_{\mathsf{e}}\delta_1}{\delta_1 - \delta_0}\right) \text{ and } \gamma = \frac{\omega}{\rho_{\mathsf{e}}}\left(\frac{\lambda_{\mathsf{e}}^{\mathsf{u}} - \rho_{\mathsf{e}}\delta_0}{\delta_0}\right), \tag{2}$$

where $\delta_n = \sigma_n(c)/\sigma_{n+1}(c)$, $\sigma_0(c) = \rho_{\mathsf{e}}^c/c!$, $\sigma_j(c) = \sum_{i=0}^{c} C_{j+i-1}^{i} \rho_{\mathsf{e}}^{c-i}/(c-i)!$, and $C_y^x = y!/(x!(y-x)!)$.

2) Blocking Probability Analysis in BWP L_{u}: We denote n_1 and n_2 as the numbers of eMBB and URLLC flows being served in BWP L_{u}, respectively and z as the state of the random switch taking on the value of 1 or 0 depending on whether the IPP is *ON* or *OFF*. We define a three-dimension Markov chain with the stage space described as follows:

$$\mathcal{S}_{\mathsf{u}} = \{\mathbf{x} = (n_1, n_2, z) | 0 \leq n_1 \leq K_{\mathsf{u}}, 0 \leq n_2 \leq N_{\mathsf{u}} - n_1, z = 1 \text{ or } 0\}. \tag{3}$$

The transition probabilities between different states are shown in Fig. 2 based on which the associated state equations can be expressed as in (4), as shown at the bottom of the next page. In (4), $p_{\mathsf{u}}(\mathbf{x})$ represents the stationary probability of $\mathbf{x}$ and $\mathbb{1}_{\mathcal{Y}}$ denotes the indicator function of the event $\mathcal{Y}$.

a) Iterative algorithm to determine stationary probabilities: Using the results in (4), as shown at the bottom of the next page, the stationary probabilities, $p_{\mathsf{u}}(\mathbf{x})$'s, can be determined by employing the Gauss–Seidel iterative algorithm [15] as described below. Let $\mathbf{p}_0$ and $\mathbf{p}_1$ be the vector representing all stationary probabilities of the feasible states when $z = 0$ and 1, respectively. Then, from (4), we have

$$(\mathbf{Q}_0 + \omega\mathbf{I})\mathbf{p}_0 = \gamma\mathbf{p}_1, \text{ and } (\mathbf{Q}_1 + \gamma\mathbf{I})\mathbf{p}_1 = \omega\mathbf{p}_0, \tag{5}$$

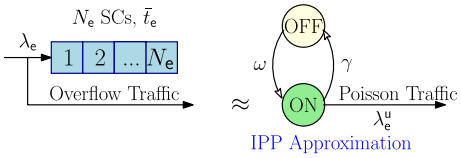

Fig. 1. The interrupted poisson process approximation.

[2]This assumption can be relaxed where a fixed number of SCs (larger than 1) is assigned to serve an active flow by scaling down the maximum number of active flows that can be served for a given number of SCs.

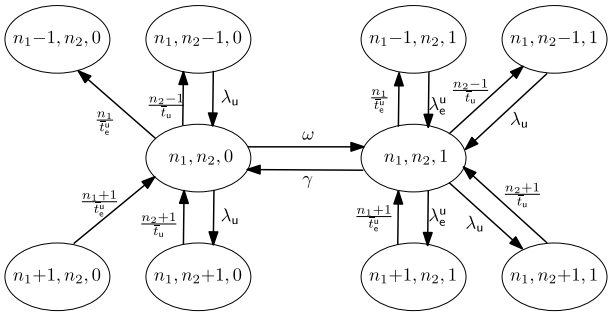

Fig. 2. The state transition diagram of three-dimension Markov chain in S_{u}.

Algorithm 1 Iterative Algorithm

1: Initialize: Choose $\mathbf{p}_0^{(0)} \neq \mathbf{0}$, a convergence criterion ε, and set $\ell = 0$.
2: **repeat**
3: Determine $\mathbf{p}_1^{(\ell+1)} = \frac{\mathbf{Q}_0+\omega\mathbf{I}}{\gamma}\mathbf{p}_0^{(\ell)}$ and $\mathbf{p}_0^{(\ell+1)} = \frac{\mathbf{Q}_1+\gamma\mathbf{I}}{\omega}\mathbf{p}_1^{(\ell+1)}$.
4: Update $\ell = \ell + 1$.
5: **until** $\max(|\mathbf{p}_z^{(\ell+1)} - \mathbf{p}_z^{(\ell)}|_2) \leq \varepsilon$.

where $\mathbf{Q}_0$ and $\mathbf{Q}_1$ are the transition matrices described as

$$\mathbf{Q}_0 = \begin{bmatrix} \mathbf{Q}_{0,0} & \mathbf{U}_{0,0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_{0,1} & \mathbf{U}_{0,1} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \cdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{Q}_{0,K_{\mathsf{u}}} \end{bmatrix}, \tag{6}$$

$$\mathbf{Q}_1 = \begin{bmatrix} \mathbf{Q}_{1,0} & \mathbf{U}_{1,0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{L}_{1,1} & \mathbf{Q}_{1,1} & \mathbf{U}_{1,1} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \cdots & \vdots & \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{L}_{1,K_{\mathsf{u}}} & \mathbf{Q}_{1,K_{\mathsf{u}}} \end{bmatrix}, \tag{7}$$

with $\mathbf{Q}_{0,m}, \mathbf{Q}_{1,m} \in \mathbb{R}^{m^*\times m^*}$, $\mathbf{U}_{0,m}, \mathbf{U}_{1,m} \in \mathbb{R}^{m^*\times(m^*-1)}$, $\mathbf{L}_{1,m} \in \mathbb{R}^{(m^*-1)\times m^*}$, $m^* = N_{\mathsf{u}} + 1 - m$, and

$$\mathbf{Q}_{0,m} = \begin{bmatrix} \frac{m}{\bar{t}_{\mathsf{e}}^{\mathsf{u}}}+\lambda_{\mathsf{u}} & -\frac{1}{\bar{t}_{\mathsf{u}}} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \frac{m}{\bar{t}_{\mathsf{e}}^{\mathsf{u}}}+\lambda_{\mathsf{u}}+\frac{1}{\bar{t}_{\mathsf{u}}} & -\frac{2}{\bar{t}_{\mathsf{u}}} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \cdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \frac{m}{\bar{t}_{\mathsf{e}}^{\mathsf{u}}}+\frac{N_{\mathsf{u}}-m}{\bar{t}_{\mathsf{u}}} \end{bmatrix}, \tag{8}$$

$$\mathbf{Q}_{1,m} = \mathbf{Q}_{0,m} + diag(\lambda_{\mathsf{e}}^{\mathsf{u}}, \ldots, \lambda_{\mathsf{e}}^{\mathsf{u}}, 0), \text{ if } m < K_{\mathsf{u}}, \tag{9}$$

$$\mathbf{U}_{0,m} = \mathbf{U}_{1,m} = -diag(\frac{m+1}{\bar{t}_{\mathsf{e}}^{\mathsf{u}}}, \ldots, \frac{m+1}{\bar{t}_{\mathsf{e}}^{\mathsf{u}}}), \tag{10}$$

$$\mathbf{L}_{1,m} = -diag(\lambda_{\mathsf{e}}^{\mathsf{u}}, \ldots, \lambda_{\mathsf{e}}^{\mathsf{u}}). \tag{11}$$

From (5), $\mathbf{p}_0$ and $\mathbf{p}_1$ can be obtained by employing the iterative algorithm described in Algorithm 1.

b) Blocking probability: With stationary probabilities obtained from Algorithm 1, we can derive the BPs of data flows in BWP L_{u} as

$$\mathsf{Pr}_{\mathsf{u,blk}} = \sum_{m=0}^{K_{\mathsf{u}}} p_{\mathsf{u}}(m, N_{\mathsf{u}}-m, 0) + p_{\mathsf{u}}(m, N_{\mathsf{u}}-m, 1), \tag{12}$$

$$\mathsf{Pr}_{\mathsf{e,blk}}^{\mathsf{U}} = \sum_{m=0}^{K_{\mathsf{u}}} p_{\mathsf{u}}(m, N_{\mathsf{u}}-m, 1) + \sum_{n=0}^{N_{\mathsf{u}}-K_{\mathsf{u}}-1} p_{\mathsf{u}}(K_{\mathsf{u}}, n, 1), \tag{13}$$

where $\mathsf{Pr}_{\mathsf{u,blk}}$ and $\mathsf{Pr}_{\mathsf{e,blk}}^{\mathsf{U}}$ are the BPs of URLLC and eMMB flows in L_{u}, respectively. Then, the overall BP of an eMBB flow can be calculated as

$$\mathsf{Pr}_{\mathsf{e,blk}} = \mathsf{Pr}_{\mathsf{e,blk}}^{\mathsf{E}} \mathsf{Pr}_{\mathsf{e,blk}}^{\mathsf{U}}. \tag{14}$$

3) Blocking Probability Characteristics: From (12)–(14), some specific characteristics of system BPs can be stated in the following propositions.

Proposition 1: $\mathsf{Pr}_{\mathsf{e,blk}}$ decreases with the traffic intensity if $\mathsf{Pr}_{\mathsf{e,blk}}^{\mathsf{E}}$ decreases.

Proof: As can be seen in (14), the smaller value of $\mathsf{Pr}_{\mathsf{e,blk}}^{\mathsf{E}}$ results in the reduction of $\mathsf{Pr}_{\mathsf{e,blk}}$. In addition, depreciating $\mathsf{Pr}_{\mathsf{e,blk}}^{\mathsf{E}}$ also lessens the intensity of the overflow traffic which, therefore, reduces its BP in BWP L_{u}, $\mathsf{Pr}_{\mathsf{e,blk}}^{\mathsf{U}}$, and hence reducing the overall BP of an eMBB flow, $\mathsf{Pr}_{\mathsf{e,blk}}$. ■

Proposition 2: Let $\rho_{\mathsf{u}} = \bar{t}_{\mathsf{u}}\lambda_{\mathsf{u}}$ and $\rho_{\mathsf{e}}^{\mathsf{u}} = \bar{t}_{\mathsf{e}}^{\mathsf{u}}\lambda_{\mathsf{e}}^{\mathsf{u}}$. For given $\lambda_{\mathsf{e}}^{\mathsf{u}}$ and N_{u}, the bounds of $\mathsf{Pr}_{\mathsf{u,blk}}$ can be defined as follows:

$$B(\rho_{\mathsf{u}}, N_{\mathsf{u}}) \leq \mathsf{Pr}_{\mathsf{u,blk}} \leq B(\rho_{\mathsf{u}} + \rho_{\mathsf{e}}^{\mathsf{u}}, N_{\mathsf{u}}). \tag{15}$$

Proof: As can be seen, the lower bound of $\mathsf{Pr}_{\mathsf{u,blk}}$ can be obtained when there is no overflow traffic from BWP L_{e}. Hence, the BP in such a scenario can be obtained from the Erlang B formula $B(\rho_{\mathsf{u}}, N_{\mathsf{u}})$. For the upper bound, the overflow traffic exercises its strongest influence on URLLC flows when $\gamma \simeq 0$ and $\omega \simeq \infty$ and $K_{\mathsf{u}} = N_{\mathsf{u}}$. In this scenario, there are two PPs with intensity ρ_{u} and $\rho_{\mathsf{e}}^{\mathsf{u}}$ in BWP L_{u}. Thus, the upper bound can be defined as $B(\rho_{\mathsf{u}} + \rho_{\mathsf{e}}^{\mathsf{u}}, N_{\mathsf{u}})$. ■

IV. Joint Admission Control and Network Slicing Optimization

We study the joint network slicing, numerology allocation, and admission control optimization problem for BWPs L_{e} and L_{u} to minimize the BP of eMBB flows while protecting the QoS of URLLC flows. This problem can be stated as

$$(\mathcal{P}_0) \min_{\mathbf{\Omega}} \mathsf{Pr}_{\mathsf{e,blk}}(\Omega) \ \text{ s.t. } \ \mathsf{Pr}_{\mathsf{u,blk}} \leq \bar{\alpha}_{\mathsf{u}}, \tag{16a}$$

$$W_{\mathsf{e}} + W_{\mathsf{u}} \leq W_{\mathsf{total}}, \tag{16b}$$

where $\Omega = \{W_{\mathsf{e}}, W_{\mathsf{u}}, g_{\mathsf{e}}, g_{\mathsf{u}}, K_{\mathsf{u}} | g_{\mathsf{e}} \in \{1,2,3,4\}, g_{\mathsf{u}} \in \{5,6\}, 0 \leq K_{\mathsf{u}} \leq N_{\mathsf{u}}\}$, $\bar{\alpha}_{\mathsf{u}}$ is the required minimum BP of the

$$(\frac{n_1}{\bar{t}_{\mathsf{e}}^{\mathsf{u}}} + \frac{n_2}{\bar{t}_{\mathsf{u}}} + \lambda_{\mathsf{u}}\mathbb{1}_{n_1+n_2<N_{\mathsf{u}}} + \omega)p_{\mathsf{u}}(n_1,n_2,0) = \frac{n_1+1}{\bar{t}_{\mathsf{e}}^{\mathsf{u}}}p_{\mathsf{u}}(n_1+1,n_2,0)\mathbb{1}_{\substack{n_1+n_2<N_{\mathsf{u}} \\ \& \ n_1<K_{\mathsf{u}}}} + \frac{n_2+1}{\bar{t}_{\mathsf{u}}}p_{\mathsf{u}}(n_1,n_2+1,0)\mathbb{1}_{n_1+n_2<N_{\mathsf{u}}}$$
$$+ \lambda_{\mathsf{u}}p_{\mathsf{u}}(n_1,n_2-1,0)\mathbb{1}_{n_2>0} + \gamma p_{\mathsf{u}}(n_1,n_2,1) \tag{4a}$$

$$(\frac{n_1}{\bar{t}_{\mathsf{e}}^{\mathsf{u}}} + \lambda_{\mathsf{e}}^{\mathsf{u}}\mathbb{1}_{\substack{n_1+n_2<N_{\mathsf{u}} \\ \& \ n_1<K_{\mathsf{u}}}} + \frac{n_2}{\bar{t}_{\mathsf{u}}} + \lambda_{\mathsf{u}}\mathbb{1}_{n_1+n_2<N_{\mathsf{u}}} + \gamma)p_{\mathsf{u}}(n_1,n_2,1) = \frac{n_1+1}{\bar{t}_{\mathsf{e}}^{\mathsf{u}}}p_{\mathsf{u}}(n_1+1,n_2,1)\mathbb{1}_{\substack{n_1+n_2<N_{\mathsf{u}} \\ \& \ n_1<K_{\mathsf{u}}}}$$
$$+ \frac{n_2+1}{\bar{t}_{\mathsf{u}}}p_{\mathsf{u}}(n_1,n_2+1,1)\mathbb{1}_{n_1+n_2<N_{\mathsf{u}}} + \lambda_{\mathsf{e}}^{\mathsf{u}}p_{\mathsf{u}}(n_1-1,n_2,1)\mathbb{1}_{n_1>0} + \lambda_{\mathsf{u}}p_{\mathsf{u}}(n_1,n_2-1,1)\mathbb{1}_{n_2>0} + \omega p_{\mathsf{u}}(n_1,n_2,0) \tag{4b}$$

URLLC service. Problem $(\mathcal{P}_0)$ optimizes three design issues with their corresponding decision variables: network slicing ($\Omega_1 = [W_{\mathsf{e}}, W_{\mathsf{u}}]$), numerology allocation ($\Omega_2 = [g_{\mathsf{e}}, g_{\mathsf{u}}]$),[3] and admission control design (K_{u}). Joint optimization of these variables results in a challenging mixed integer programming problem. Hence, we propose to decompose this problem into several low-complexity sub-problems as follows.

1) Numerology Allocation for BWP L_{e}: Thanks to Proposition 1, $\mathsf{Pr}_{\mathsf{e,blk}}(\Omega)$ can be reduced by minimizing $\mathsf{Pr}^{\mathsf{E}}_{\mathsf{e,blk}}$. Therefore, the optimal numerology for BWP L_{e} for given N_{e} can be determined as follows:

$$g^{\star}_{\mathsf{e}} = \underset{g_{\mathsf{e}}\in\{1,2,3,4\}}{\arg\min}\ \mathsf{Pr}^{\mathsf{E}}_{\mathsf{e,blk}} = \underset{g_{\mathsf{e}}\in\{1,2,3,4\}}{\arg\min}\ B\left(\frac{\lambda_{\mathsf{e}}}{R(g_{\mathsf{e}})\mu_{\mathsf{e}}}, N_{\mathsf{e}}\right). \quad (17)$$

2) Admission Control Parameter Design: It can be verified that for given N_{u} and g_{u}, increasing K_{u} results in the decrease of $\mathsf{Pr}^{\mathsf{U}}_{\mathsf{e,blk}}$ which also lessens $\mathsf{Pr}_{\mathsf{e,blk}}(\Omega)$. Hence, the optimal value of K_{u} can be determined as

$$K^{\star}_{\mathsf{u}} = \max_{0\le K_{\mathsf{u}}\le N_{\mathsf{u}}} K_{\mathsf{u}} \text{ s.t. } \mathsf{Pr}_{\mathsf{u,blk}} \le \bar{\alpha}_{\mathsf{u}}. \quad (18)$$

3) Network Slicing Design: For given Ω_2 and K_{u}, the network slicing problem can be stated as

$$\min_{\mathbf{\Omega}_1}\ \mathsf{Pr}_{\mathsf{e,blk}}(\Omega) \text{ s.t. constraints (16a) and (16b).} \quad (19)$$

From (16b), the upper bounds of N_{e} and N_{u} can be given as

$$N_{\mathsf{e}} \le \lfloor W_{\mathsf{total}}/(W_0 2^{g_{\mathsf{e}}})\rfloor \text{ and } N_{\mathsf{u}} \le \lfloor W_{\mathsf{total}}/(W_0 2^{g_{\mathsf{u}}})\rfloor. \quad (20)$$

Let $B^{-1}(\rho,\alpha)$ be the inverse function of $B(\rho,n)$, i.e., $B^{-1}(\rho,\alpha) = \min_n n \text{ s.t. } B(\rho,n)\le\alpha$. It is worth noting that $B(\rho,n)$ is a monotonic decreasing function with respect to n for given ρ; hence, $B^{-1}_n(\rho,\alpha)$ is unique for given ρ and α.

Proposition 3: For given g_{u}, the required bandwidth of L_{u} can be bounded as

$$B^{-1}_n(\rho_{\mathsf{u}},\bar{\alpha}_{\mathsf{u}}) \le N_{\mathsf{u}} = \frac{W_{\mathsf{u}}}{W_0 2^{g_{\mathsf{u}}}} \le B^{-1}_n\left(\frac{\lambda_{\mathsf{u}}\mu_{\mathsf{e}}+\lambda^{\mathsf{u}}_{\mathsf{e}}\mu_{\mathsf{u}}}{R(g_{\mathsf{u}})\mu_{\mathsf{u}}\mu_{\mathsf{e}}}, \bar{\alpha}_{\mathsf{u}}\right). \quad (21)$$

Proof: It can be verified that the lower and upper bounds given in this proposition can be obtained directly from Proposition 2 to satisfy the constraint (16a) and $\rho_{\mathsf{u}} + \rho^{\mathsf{u}}_{\mathsf{e}} = \frac{\lambda_{\mathsf{u}}\mu_{\mathsf{e}}+\lambda^{\mathsf{u}}_{\mathsf{e}}\mu_{\mathsf{u}}}{R(g_{\mathsf{u}})\mu_{\mathsf{u}}\mu_{\mathsf{e}}}$. ■

The result given in Proposition 3 enables us reduce the research range for N_{u} from $[0, W_{\mathsf{total}}2^{-g_{\mathsf{u}}}/0.18]$ to

$$\mathcal{N}_{\mathsf{u}} = \left\{N\in\mathcal{N} | B^{-1}_n(\rho_{\mathsf{u}},\bar{\alpha}_{\mathsf{u}}) \le N \le B^{-1}_n(\rho_{\mathsf{u,e}},\bar{\alpha}_{\mathsf{u}})\right\}, \quad (22)$$

where $\rho_{\mathsf{u,e}} = \rho_{\mathsf{u}} + \rho^{\mathsf{u}}_{\mathsf{e}}$. Additionally, for a given N_{u}, the optimal value of N_{e} can be expressed as

$$N_{\mathsf{e}} = \lfloor (W_{\mathsf{total}} - W_0 N_{\mathsf{u}} 2^{g_{\mathsf{u}}})/(W_0 2^{g_{\mathsf{e}}})\rfloor, \quad (23)$$

because the larger bandwidth is allocated for L_{e}, the smaller values of $\mathsf{Pr}^{\mathsf{E}}_{\mathsf{e,blk}}$ and $\mathsf{Pr}_{\mathsf{e,blk}}$ that can be achieved. The required range of W_{u} becomes smaller once N_{e} is updated so that the intensity of the overflow traffic decreases. Thanks to this observation, the optimal value of $\mathbf{\Omega}_1$ can be obtained by iteratively updating $\mathcal{N}_{\mathsf{u}}$ and N_{e}, based on which we propose an efficient searching algorithm as summarized in Algorithm 2 to obtain the optimal solution of $(\mathcal{P}_0)$.

[3]According the 5G NR standard, one unique numerology should be allocated for each bandwidth part [11].

Algorithm 2 Proposed Searching Algorithm

1: Initialize: Set $\mathsf{Pr}^{\star}_{\mathsf{e,blk}} = 1$.
2: **for** $g_{\mathsf{u}} \in \{5,6\}$ **do**
3: Set $\lambda^{\mathsf{u}}_{\mathsf{e}} = \lambda_{\mathsf{e}}$.
4: **repeat**
5: Choose $N_{\mathsf{u}} = B^{-1}_n(\rho_{\mathsf{u,e}},\bar{\alpha}_{\mathsf{u}})$.
6: Update N_{e} as in (23), g_{e} as in (17), and $\lambda^{\mathsf{u}}_{\mathsf{e}}$ as in (2).
7: **until** N_{e} is unchanged.
8: Determine $\mathcal{N}_{\mathsf{u}}$ as in (22).
9: **for** $N_{\mathsf{u}} \in \mathcal{N}_{\mathsf{u}}$ **do**
10: Update N_{e} as in (23) and $(\lambda^{\mathsf{u}}_{\mathsf{e}},\omega,\gamma)$ as in (2).
11: Determine $g_{\mathsf{e}}, K_{\mathsf{u}}$ as in (17),(18) and calculate $\mathsf{Pr}_{\mathsf{e,blk}}(\Omega)$.
12: **if** $\mathsf{Pr}_{\mathsf{e,blk}}(\Omega) \le \mathsf{Pr}^{\star}_{\mathsf{e,blk}}$ **then**
13: Set $\Omega^{\star} = \Omega$ and $\mathsf{Pr}^{\star}_{\mathsf{e,blk}} = \mathsf{Pr}_{\mathsf{e,blk}}(\Omega)$.
14: **end if**
15: **end for**
16: **end for**
17: Return $\Omega^{\star}$.

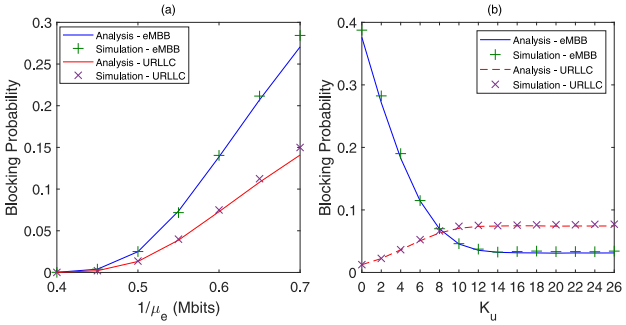

Fig. 3. BPs of eMBB and URLLC traffic vs $1/\mu_{\mathsf{u}}$ (a) and K_{u} (b).

V. Numerical Results

In this section, we first validate the accuracy of the proposed analytical framework presented in Section III via simulation, then, we study the efficiency of Algorithm 2 under different parameter settings. In the simulation, we consider the 5G-NR wireless system operating at the frequency band 3.6−3.8 GHz, i.e., $W_{\mathsf{total}} = 200$ MHz [16]. Assume that the 16-QAM modulation scheme is employed by all traffic flows, i.e., $r_{\mathsf{Hz}} = 4$ bits/Hz [11]. Unless stated otherwise, the parameters are set as follows: $\lambda_{\mathsf{e}} = \lambda_{\mathsf{u}} = 40$, $1/\mu_{\mathsf{e}} = 1/\mu_{\mathsf{u}} = 1$ Mbits, $\bar{\alpha}_u = 10^{-3}$. To obtain simulation results for some specific values of $\lambda_{\mathsf{e}}, \lambda_{\mathsf{u}}, \mu_{\mathsf{e}}, \mu_{\mathsf{u}}$ and Ω, we generate over 10^8 traffic flow samples for the eMBB and URLLC services following the corresponding Poisson processes as follows. The arrival time of a new flow is determined based on the arrival time of its immediately preceding flow and a random inter-arrival time that is generated from MATLAB according to an exponential distribution with the corresponding values of $\lambda_{\mathsf{e}}, \lambda_{\mathsf{u}}$. Similarly, the data length brought by a new flow is also generated randomly according to the exponential distribution with the corresponding values of $\mu_{\mathsf{e}}, \mu_{\mathsf{u}}$. The obtained data length is then used to estimate the completed transmission time. A traffic flow is assumed to occupy one sub-channel during the interval between the flow's arrival instant and the completed transmission instant. Then, the bounding admission control strategy described in Section II is implemented and the number of blocked flows is counted during the simulation based on which we calculate the blocking probability.

Fig. 3 illustrates the BPs of eMMB and URLLC traffic, obtained by the proposed analytical framework and simulation for different values of $1/\mu_{\mathsf{u}}$ and K_{u}, respectively. As

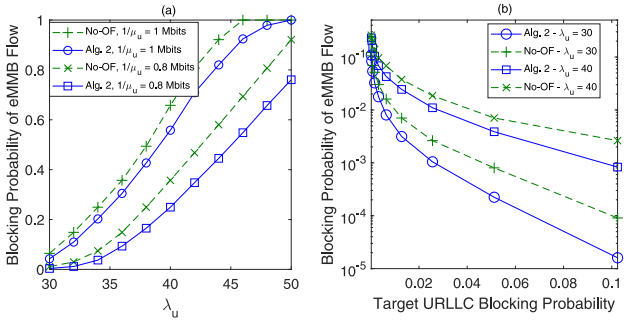

Fig. 4. BP of eMBB traffic vs λ_u (a) and $\bar{\alpha}_\mathsf{u}$ (b).

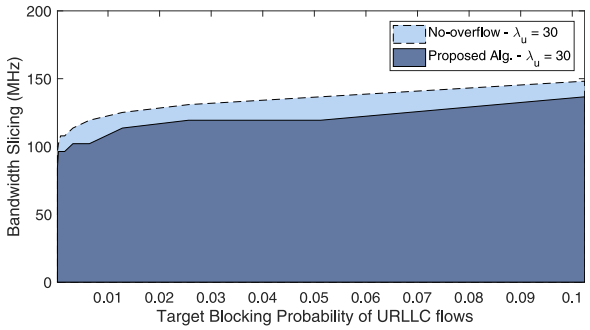

Fig. 5. Bandwidth allocated for eMBB traffic vs $\bar{\alpha}_\mathsf{u}$.

can be seen, results achieved by the proposed analysis are in good agreement with the simulation results which confirms the accuracy of the IPP approximation. With $K_\mathsf{u} = N_\mathsf{u}$, Fig. 3-(a) shows that increasing the average data length of the URLLC flow results in higher BPs for both services. In addition, Fig. 3-(b) demonstrates that adopting the overflow strategy can help mitigate the overload of the eMBB BWP but it degrades the BP of the URLLC service. Interestingly, both BPs saturate when K_u becomes sufficiently large.

Fig. 4 shows the BP of eMMB traffic versus the arrival rate λ_u and target blocking probability of URLLC traffic $\bar{\alpha}_\mathsf{u}$ for two schemes, namely our optimized design using Algorithm 2 with overflow and the conventional scheme with no-overflow (indicated as "Alg. 2" and "No-OF" in these figures, respectively). For the no-overflow scheme, the minimum value of W_u is first determined so that $B(\rho_\mathsf{u}, N_\mathsf{u}) \leq \bar{\alpha}_\mathsf{u}$. Then, the remaining bandwidth is allocated to serve the eMMB traffic. The numerology is also optimized in each BWP. As can be observed, the $\mathsf{Pr}_{\mathsf{e,blk}}$ achieved by our proposed design is much lower than that due to the no-overflow scheme. In addition, the $\mathsf{Pr}_{\mathsf{e,blk}}$ achieved by both schemes increase as λ_u increases and decrease as $\bar{\alpha}_\mathsf{u}$ increases. This happens because the higher arrival rate of URLLC traffic and its lower target BP both result in higher traffic load at BWP L_u. This may degrade the performance of eMBB data transmission.

The network slicing result is illustrated in Fig. 5 in which the bandwidth of the BWP assigned to eMBB service due to the proposed algorithm and the no-overflow scheme is plotted versus $\bar{\alpha}_\mathsf{u}$. Interestingly, our proposed framework allocates less bandwidth to the eMBB traffic compared to that due to the no-overflow scheme, but it delivers better performance, which again confirms the benefit of adopting our design framework.

VI. Conclusion

In this letter, we have proposed a novel design framework for joint network slicing, numerology allocation, and admission control to support the eMBB and URLLC services. Using the bounding admission control strategy, the BPs of eMBB and URLLC traffic have been analyzed and an efficient searching algorithm has been proposed to minimize the BP of eMBB traffic while maintaining the BP requirements of URLLC traffic. Numerical results have confirmed the accuracy of the proposed analytical framework and the benefit of employing the overflow strategy in the admission control.

References

[1] S.-Y. Lien, S.-L. Shieh, Y. Huang, B. Su, Y.-L. Hsu, and H.-Y. Wei, "5G new radio: Waveform, frame structure, multiple access, and initial access," *IEEE Commun. Mag.*, vol. 55, no. 6, pp. 64–71, Jun. 2017.

[2] E. Fountoulakis, N. Pappas, Q. Liao, V. Suryaprakash, and D. Yuan, "An examination of the benefits of scalable TTI for heterogeneous traffic management in 5G networks," in *Proc. IEEE Int. Symp. Model. Optim. Mobile Ad Hoc Wireless Netw. (WiOpt)*, 2017, pp. 1–6.

[3] T. T. Nguyen, V. N. Ha, and L. B. Le, "Wireless scheduling for heterogeneous services with mixed numerology in 5G wireless networks," *IEEE Commun. Lett.*, to be published.

[4] 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. 5565–55779, 2018.

[5] J. Park and M. Bennis, "URLLC-eMBB slicing to support VR multimodal perceptions over wireless cellular systems," in *Proc. IEEE Glob. Commun. Conf. (GLOBECOM)*, 2018, pp. 1–7.

[6] A. Anand, G. De Veciana, and S. Shakkottai, "Joint scheduling of URLLC and eMBB traffic in 5G wireless networks," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, 2018, pp. 1970–1978.

[7] A. A. Esswie and K. I. Pedersen, "Opportunistic spatial preemptive scheduling for URLLC and eMBB coexistence in multi-user 5G networks," *IEEE Access*, vol. 6, pp. 38451–38463, 2018.

[8] Y. Fang and Y. Zhang, "Call admission control schemes and performance analysis in wireless mobile networks," *IEEE Trans. Veh. Technol.*, vol. 51, no. 2, pp. 371–382, Mar. 2002.

[9] M. Maternia *et al.*, "5G PPP use cases and performance evaluation models, version 1.0," Valencia, Spain, 5G PPP, white paper, Apr. 2016. Accessed: Oct. 17, 2019. [Online]. Available: https://5g-ppp.eu/wp-content/uploads/2014/02/5G-PPP-use-cases-and-performance-evaluation-modeling_v1.0.pdf

[10] A. Benjebbour, K. Kitao, Y. Kakishima, and C. Na, "3GPP defined 5G requirements and evaluation conditions," *NTT DOCOMO Tech. J.*, vol. 19, no. 3, pp. 13–23, Jan. 2018.

[11] "NR; Physical channels and modulation, V15.6.0," Standard TS 38.211, Jun. 2019. Accessed: Aug. 23, 2019. [Online]. Available: https://www.3gpp.org/ftp/Specs/archive/38_series/38.211/38211–f60.zip

[12] J. L. Davis, W. A. Massey, and W. Whitt, "Sensitivity to the service-time distribution in the nonstationary Erlang loss model," *Manag. Sci.*, vol. 41, no. 6, pp. 1107–1116, Jun. 1995.

[13] Y.-C. Chan and E. W. M. Wong, "Blocking probability evaluation for non-hierarchical overflow loss systems," *IEEE Trans. Commun.*, vol. 66, no. 5, pp. 2022–2036, May 2018.

[14] A. Kuczura, "The interrupted Poisson process as an overflow process," *Bell Syst. Tech. J.*, vol. 52, no. 3, pp. 437–448, Mar. 1973.

[15] D. Gross and C. M. Harris, *Fundamentals of Queueing Theory*, 3rd ed. New York, NY, USA: Wiley, 1998.

[16] *European 5G Observatory*. Accessed: Oct. 23, 2019. [Online]. Available: https://5gobservatory.eu/5g-spectrum/national-5g-spectrum-assignment/