Joint Resource Block Allocation and Beamforming with Mixed-Numerology for eMBB and URLLC Use Cases

Zihuan Wang and Vincent W.S. Wong
Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, Canada
email: {zihuanwang, vincentw}@ece.ubc.ca

***Abstract*—Mixed-numerology has been proposed in the Third Generation Partnership Project (3GPP) standard for the fifth generation (5G) wireless networks, where flexible subcarrier spacing (SCS) can be applied to support uses cases with different quality-of-service (QoS) requirements. In this paper, we study the joint design of resource block allocation and beamforming with mixed-numerology for enhanced mobile broadband (eMBB) and ultra-reliable low-latency communications (URLLC) use cases. We consider multiple multi-antenna base stations (BSs) cooperatively provide services to the users. By using beamforming, inter-user interference can be mitigated and a resource block can be utilized by more than one user. Short packet transmission is considered for URLLC users to satisfy their low-latency requirements. We formulate a mixed-integer nonlinear programming problem to maximize the aggregate throughput of eMBB users while guaranteeing the throughput, reliability, and latency requirements of URLLC users. We propose a low-complexity algorithm, which leverages fractional programming and successive convex approximation (SCA), to obtain the solutions. Simulation results show that our proposed algorithm can improve the aggregate eMBB throughput by 30% compared with the fixed-numerology based approach.**

I. Introduction

Wireless networks are evolving quickly due to the ever-increasing traffic and growing number of devices. The fifth generation (5G) wireless networks are targeting to support various applications with diverse quality of service (QoS) requirements of throughput, latency, and reliability. Three use cases have been identified in 5G networks, which are the enhanced mobile broadband (eMBB), ultra-reliable and low-latency communications (URLLC), and massive machine-type communications (mMTC) [1].

Different approaches have been proposed in the literature to support eMBB and URLLC use cases in 5G networks (e.g., [2]-[4]). Network slicing has been utilized to dynamically allocate resources to eMBB and URLLC slices [2] and [3]. In order to achieve low-latency of URLLC services, one approach is to reduce the number of symbols to be transmitted to URLLC users, e.g., by using mini-slots transmission instead of the whole transmission time interval (TTI). The URLLC and eMBB services can then be multiplexed through preemption and puncturing [4]. However, with puncturing, transmission of URLLC packets may occupy the resources for eMBB services. This approach may lead to a throughput reduction of eMBB services.

Recently, mixed-numerology, which supports various subcarrier spacings (SCSs), has been proposed and standardized in the Third Generation Partnership Project (3GPP) standardization body [5]. In this way, different types of use cases have different TTIs and SCSs. A numerology refers to SCS and cyclic prefix (CP) overhead. In Long Term Evolution (LTE) systems, fixed-numerology of 15 kHz SCS and 1.0 ms TTI is utilized and applied to all resource blocks. The 5G New Radio (NR) uses mixed-numerology and supports scalable symbol durations. The SCS values are defined as $15 \times 2^{\mu}$ kHz, where $\mu \in \{0, 1, 2, 3, 4, 5\}$. The symbol length, including CP of 15 kHz SCS, is equal to the sum of 2^{μ} symbols of the $15 \times 2^{\mu}$ kHz SCS. Then, the QoS requirements of different use cases can be satisfied by using proper numerologies.

The design of resource block allocation using mixed-numerology is crucial when different types of use cases coexist in 5G networks. In [6], an iterative resource allocation algorithm based on Lagrange duality is proposed to maximize the total throughput of different type of services. In [7], an integer linear programming problem is formulated to maximize the number of scheduled users. The problem is solved by using resource partitioning and iterative greedy algorithms. A deep Q-learning algorithm is proposed in [8] to support flexible numerology. In [9], a joint transmission power and resource block allocation scheme with mixed-numerology is proposed to minimize the transmit power at the base station (BS) while satisfying different QoS requirements. In [10], an algorithm to support eMBB and URLLC services with mixed-numerology is proposed to maximize the energy efficiency.

The aforementioned works [6]-[10] consider single-input single-output (SISO) communications, and each resource block is only assigned to one user. In this paper, we consider multiple multi-antenna BSs cooperatively provide services to users through beamforming. The use of beamforming can improve the throughput and enable the resource blocks to be utilized by multiple users. The contributions of this paper are summarized as follows:

- We consider a wireless communications system with multiple BSs cooperatively serving eMBB and URLLC users by using different numerologies. Short packet transmission is considered for URLLC use cases. We aim to maximize the aggregate throughput of eMBB users while guaranteeing the throughput, reliability, and latency

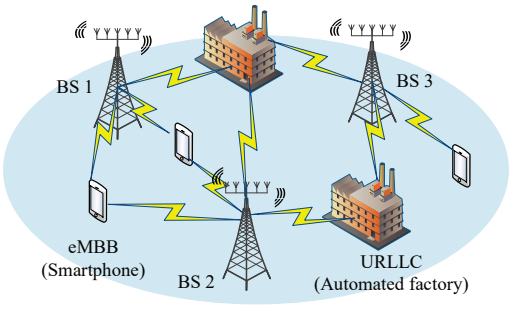

Fig. 1. A wireless system with three BSs cooperatively serving eMBB and URLLC users.

requirements of URLLC users.

- We formulate the problem as a mixed-integer nonlinear programming problem. The formulated problem is non-convex and difficult to solve due to the consideration of short packet transmission for URLLC users and the inter-user interference in the objective function. To tackle these issues, we propose a suboptimal solution with polynomial computational complexity, which leverages fractional programming and successive convex approximation (SCA) techniques.
- Simulation results show that our proposed algorithm with mixed-numerology can improve the aggregate throughput of eMBB users by 30% when compared with the fixed-numerology based approach. Moreover, when the URLLC latency becomes more stringent, our proposed algorithm results in a lower throughput reduction than the fixed-numerology approach.

The rest of this paper is organized as follows. The system model and problem formulation are presented in Section II. In Section III, we present the proposed algorithm based on fractional programming and SCA to solve the formulated problem. Performance evaluation is presented in Section IV. Finally, conclusions are drawn in Section V.

In this paper, we use boldface lower case letters and boldface upper case letters to denote vectors and matrices, respectively. $(\cdot)^*$, $(\cdot)^T$, and $(\cdot)^H$ are used to denote the conjugate, transpose, and conjugate transpose operations, respectively. $\mathbb{C}^N$ denotes the set of N dimensional vectors with complex entries. We use $|\mathcal{A}|$ to denote the cardinality of set $\mathcal{A}$. $Re\{\cdot\}$ is used to extract the real part of a complex number.

II. System Model and Problem Formulation

We consider downlink transmission in a wireless communications system, where multiple BSs cooperatively serve the eMBB users and URLLC users, as shown in Fig. 1. Each BS is equipped with N_t antennas and each user has a single antenna. Let $\mathcal{M}$, $\mathcal{K}_\mathrm{e}$, and $\mathcal{K}_\mathrm{u}$ denote the sets of BSs, eMBB users, and URLLC users, respectively. We use $\mathcal{K}$ to denote the set of all users. We have $\mathcal{K} = \mathcal{K}_\mathrm{e} \cup \mathcal{K}_\mathrm{u}$ and $\mathcal{K}_\mathrm{e} \cap \mathcal{K}_\mathrm{u} = \emptyset$. Let $M = |\mathcal{M}|$, $K_\mathrm{e} = |\mathcal{K}_\mathrm{e}|$, and $K_\mathrm{u} = |\mathcal{K}_\mathrm{u}|$ denote the number of BSs, number of eMBB users, and number of URLLC users, respectively.

The BSs serve the users by using beamforming based on the time-frequency resources with mixed-numerology. We consider three types of numerologies, i.e., numerology-0 with

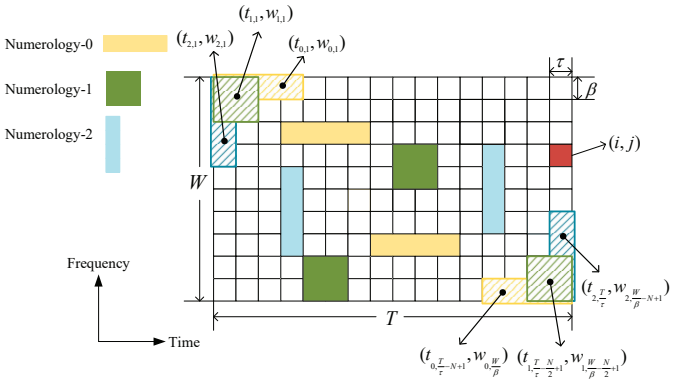

Fig. 2. Resource block allocation with mixed-numerology.

SCS of 15 kHz, numerology-1 with SCS of 30 kHz, and numerology-2 with SCS of 60 kHz. Fig. 2 shows a resource grid with total bandwidth of W and time duration of T. Each resource element has a bandwidth of $\beta = 180$ kHz in the frequency axis and $\tau = 0.125$ ms in the time axis. Each resource block contains 12 consecutive subcarriers. The bandwidths of resource block for numerology-0, 1, and 2 are given by 180 kHz, 360 kHz, and 720 kHz, respectively. Their corresponding TTI durations are 0.5 ms, 0.25 ms, and 0.125 ms, respectively. In this case, each resource block contains $N = 4$ resource elements. Let $\mathcal{T}_0 = \{t_{0,1}, \ldots, t_{0,\frac{T}{\tau}-N+1}\}$ and $\mathcal{W}_0 = \{w_{0,1}, \ldots, w_{0,\frac{W}{\beta}}\}$ denote the time and frequency index sets of the resource blocks using numerology-0 (i.e., yellow rectangles in Fig. 2). Note that $\frac{T}{\tau} - N + 1$ and $\frac{W}{\beta}$ are the total number of indices in the time axis and frequency axis for numerology-0, respectively. Similarly, let $\mathcal{T}_1 = \{t_{1,1}, \ldots, t_{1,\frac{T}{\tau}-\frac{N}{2}+1}\}$ and $\mathcal{W}_1 = \{w_{1,1}, \ldots, w_{1,\frac{W}{\beta}-\frac{N}{2}+1}\}$ denote the index sets of time-frequency resource blocks using numerology-1 (i.e., green rectangles in Fig. 2). Let $\mathcal{T}_2 = \{t_{2,1}, \ldots, t_{2,\frac{T}{\tau}}\}$ and $\mathcal{W}_2 = \{w_{2,1}, \ldots, w_{2,\frac{W}{\beta}-N+1}\}$ denote the index sets of time-frequency resource blocks using numerology-2 (i.e., blue rectangles in Fig. 2). We define $\mathcal{T}$ and $\mathcal{W}$ as the sets of all resource blocks in the time-frequency dimension. We have $\mathcal{T} = \mathcal{T}_0 \cup \mathcal{T}_1 \cup \mathcal{T}_2$ and $\mathcal{W} = \mathcal{W}_0 \cup \mathcal{W}_1 \cup \mathcal{W}_2$.

For $t \in \mathcal{T}$ and $w \in \mathcal{W}$, we introduce a binary variable $x^{(t,w)} \in \{0, 1\}$ to indicate whether resource block (t, w) is utilized. If resource block (t, w) is being used, then $x^{(t,w)}$ is equal to 1. Otherwise $x^{(t,w)}$ is equal to 0. We use $\mathcal{I} = \{(i, j) \mid i = 1, \ldots, T/\tau, j = 1, \ldots, W/\beta\}$ to denote the set of all resource elements. To map each resource block (t, w) with its resource element (i, j), we introduce binary constant $y_{(i,j)}^{(t,w)}$, which is equal to 1 if resource block (t, w) includes resource element $(i, j) \in \mathcal{I}$. Otherwise $y_{(i,j)}^{(t,w)} = 0$. For example in Fig. 2, for resource block $(t_{0,1}, w_{0,1})$ which includes resource elements $\{(1, 1), (1, 2), (1, 3), (1, 4)\}$, we have $y_{(1,1)}^{(t_{0,1},w_{0,1})} = y_{(1,2)}^{(t_{0,1},w_{0,1})} = y_{(1,3)}^{(t_{0,1},w_{0,1})} = y_{(1,4)}^{(t_{0,1},w_{0,1})} = 1$. For other elements $(i, j) \in \mathcal{I} \setminus \{(1, 1), (1, 2), (1, 3), (1, 4)\}$, we have $y_{(i,j)}^{(t_{0,1},w_{0,1})} = 0$. Note that some resource blocks may contain the same resource elements (e.g., resource blocks $(t_{0,1}, w_{0,1})$, $(t_{1,1}, w_{1,1})$, $(t_{2,1}, w_{2,1})$). To ensure that the allocated resource blocks do not overlap with each other, i.e., each

resource element is not occupied more than once, we require $\sum_{t\in\mathcal{T}}\sum_{w\in\mathcal{W}} x^{(t,w)} y_{(i,j)}^{(t,w)} \leq 1,\ (i,j)\in\mathcal{I}$.

The channel gain between user k and BS m using resource block (t,w) is denoted as $\boldsymbol{h}_{k,m}^{(t,w)} \in \mathbb{C}^{N_\mathrm{t}}$. Let $\boldsymbol{v}_{k,m}^{(t,w)} \in \mathbb{C}^{N_\mathrm{t}}$ denote the beamforming vector from BS m to user k using resource block (t,w). The signal-to-interference-plus-noise ratio (SINR) of user k on resource block (t,w) is given by

$$\gamma_k^{(t,w)} = \frac{|\sum_{m=1}^{M}(\boldsymbol{h}_{k,m}^{(t,w)})^H \boldsymbol{v}_{k,m}^{(t,w)}|^2}{\sum_{l\in\mathcal{K}\setminus\{k\}}|\sum_{m=1}^{M}(\boldsymbol{h}_{k,m}^{(t,w)})^H \boldsymbol{v}_{l,m}^{(t,w)}|^2 + (\sigma_k^{(t,w)})^2}, \quad k\in\mathcal{K}, t\in\mathcal{T}, w\in\mathcal{W}, \quad (1)$$

where $(\sigma_k^{(t,w)})^2$ represents the noise power of user k using resource block (t,w).

For eMBB users, the achievable throughput of user $k\in\mathcal{K}_\mathrm{e}$ can be expressed as follows:

$$R_{\mathrm{e},k} = \sum_{t\in\mathcal{T}}\sum_{w\in\mathcal{W}} x^{(t,w)}\beta^{(t,w)}\tau^{(t,w)}\log_2\left(1+\gamma_k^{(t,w)}\right), \quad (2)$$

where $\beta^{(t,w)}$ and $\tau^{(t,w)}$ are the bandwidth and time duration of resource block (t,w), respectively.

For URLLC service, packets are typically very short in order to achieve low latency, which also makes transmission errors unavoidable. To capture the reliability and throughput performance of URLLC users, we consider the finite blocklength coding [11]. Let L denote the length of a codeword block. The throughput of user $k\in\mathcal{K}_\mathrm{u}$ is given by

$$R_{\mathrm{u},k} = \sum_{t\in\mathcal{T}}\sum_{w\in\mathcal{W}} x^{(t,w)}\beta^{(t,w)}\tau^{(t,w)}\left(\log_2\left(1+\gamma_k^{(t,w)}\right) - \frac{Q^{-1}(\epsilon)\log_2 e}{\sqrt{L}} G_k^{(t,w)}\right), \quad (3)$$

where $Q^{-1}(\cdot)$ is the inverse of the Gaussian Q-function and ϵ is the transmission error probability. $G_k^{(t,w)}$ is the channel dispersion of user k on resource block (t,w) and is given by:

$$G_k^{(t,w)} = \sqrt{1-(1+\gamma_k^{(t,w)})^{-2}},\ k\in\mathcal{K}, t\in\mathcal{T}, w\in\mathcal{W}. \quad (4)$$

In this paper, we aim to maximize the aggregate throughput of eMBB users while guaranteeing the throughput, reliability, and latency of URLLC users. The problem can be formulated as follows:

$$\underset{\substack{x^{(t,w)},\boldsymbol{v}_{k,m}^{(t,w)}\\ k\in\mathcal{K},m\in\mathcal{M},t\in\mathcal{T},w\in\mathcal{W}}}{\text{maximize}} \quad \sum_{k\in\mathcal{K}_\mathrm{e}} \alpha_{\mathrm{e},k} R_{\mathrm{e},k} \quad (5a)$$

$$\text{subject to} \quad x^{(t,w)}\in\{0,1\},\ t\in\mathcal{T}, w\in\mathcal{W} \quad (5b)$$

$$\sum_{t\in\mathcal{T}}\sum_{w\in\mathcal{W}} x^{(t,w)} y_{(i,j)}^{(t,w)} \leq 1,\ (i,j)\in\mathcal{I} \quad (5c)$$

$$\|\boldsymbol{v}_{k,m}^{(t,w)}\| = 0,\ \forall t > T^{\max}/\tau - N + 1, t\in\mathcal{T}_0, \forall t > T^{\max}/\tau - N/2 + 1, t\in\mathcal{T}_1, \forall t > T^{\max}/\tau, t\in\mathcal{T}_2, k\in\mathcal{K}_\mathrm{u}, w\in\mathcal{W}, m\in\mathcal{M} \quad (5d)$$

$$R_{\mathrm{u},k} \geq R^{\min},\ k\in\mathcal{K}_\mathrm{u} \quad (5e)$$

TABLE I
SUMMARY OF NOTATIONS

Symbol	Definition
$\mathcal{M}$ / M	Set of BSs / Number of BSs
$\mathcal{K}_\mathrm{e}$ / K_e	Set of eMBB users / Number of eMBB users
$\mathcal{K}_\mathrm{u}$ / K_u	Set of URLLC users / Number of URLLC users
$\mathcal{K}$	Set of all the users
N_t	Number of antennas at each BS
T	Total time duration
W	Total bandwidth
τ	Time duration of a resource element
β	Bandwidth of a resource element
N	Number of resource elements in each resource block
L	Length of a codeword block
$\mathcal{T}$	Set of time index of resource block
$\mathcal{W}$	Set of band index of resource block
$R^{\min}$	Minimum throughput requirement of URLLC users
$P^{\max}$	Maximum power at each BS
$T^{\max}$	Maximum tolerated latency of URLLC users

$$\sum_{k\in\mathcal{K}}\sum_{t\in\mathcal{T}}\sum_{w\in\mathcal{W}} \|\boldsymbol{v}_{k,m}^{(t,w)}\|^2 \leq P^{\max},\ m\in\mathcal{M}. \quad (5f)$$

Constraint (5c) guarantees that the allocated resource blocks should not overlap with each other. Constraint (5d) is the latency requirement of URLLC users. The URLLC data packets for user k should be transmitted within $T^{\max}$ duration. Constraint (5e) represents the minimum throughput of URLLC users. Note that this throughput requirement will also guarantee the reliability (i.e., transmission error probability) of URLLC users [12]. Constraint (5f) corresponds to the total power constraint at each BS. The definitions of notations are summarized in Table I.

Problem (5) is a nonconvex problem due to the binary constraint (5b), as well as the non-convexity of the SINR expression in $R_{\mathrm{e},k}$ and $R_{\mathrm{u},k}$. In the next section, we propose a suboptimal algorithm with polynomial computational complexity, which uses fractional programming and SCA approaches, to solve the formulated problem.

III. Proposed Algorithm

For simplicity of representation, we define beamforming vector $\boldsymbol{v}_k^{(t,w)} \in \mathbb{C}^{MN_\mathrm{t}}$ for user k using resource block (t,w) as

$$\boldsymbol{v}_k^{(t,w)} = [(\boldsymbol{v}_{k,1}^{(t,w)})^T \cdots (\boldsymbol{v}_{k,m}^{(t,w)})^T \cdots (\boldsymbol{v}_{k,M}^{(t,w)})^T]^T, \quad (6)$$

which combines the beamformer from all BSs to user k. Similarly, we define channel vector $\boldsymbol{h}_k^{(t,w)} \in \mathbb{C}^{MN_\mathrm{t}}$ as

$$\boldsymbol{h}_k^{(t,w)} = [(\boldsymbol{h}_{k,1}^{(t,w)})^T \cdots (\boldsymbol{h}_{k,m}^{(t,w)})^T \cdots (\boldsymbol{h}_{k,M}^{(t,w)})^T]^T. \quad (7)$$

Then, the SINR expression in (1) can be rewritten as

$$\gamma_k^{(t,w)} = \frac{|(\boldsymbol{h}_k^{(t,w)})^H \boldsymbol{v}_k^{(t,w)}|^2}{\sum_{l\in\mathcal{K}\setminus\{k\}}|(\boldsymbol{h}_k^{(t,w)})^H \boldsymbol{v}_l^{(t,w)}|^2 + (\sigma_k^{(t,w)})^2}. \quad (8)$$

Notice that in the objective function (5a) and constraint (5e), the throughput expression contains the product operation between variables $x^{(t,w)}$ and $\boldsymbol{v}_k^{(t,w)}$, which makes the problem

to be nonconvex. To tackle this issue, we introduce the following constraints:

$$0 \leq \|\boldsymbol{E}_m \boldsymbol{v}_k^{(t,w)}\|^2 \leq x^{(t,w)} P^{\max}, \quad k \in \mathcal{K}, m \in \mathcal{M}, t \in \mathcal{T}, w \in \mathcal{W}, \quad (9)$$

where matrix $\boldsymbol{E}_m$ of dimension $N_{\mathrm{t}} \times N_{\mathrm{t}} M$ is a shaping matrix. It extracts the m-th subvector in $\boldsymbol{v}_k^{(t,w)}$, i.e., the beamformer from BS m. By introducing constraint (9), problem (5) can be reformulated as:

$$\underset{\substack{x^{(t,w)}, \boldsymbol{v}_k^{(t,w)} \\ k \in \mathcal{K}, t \in \mathcal{T}, w \in \mathcal{W}}}{\text{maximize}} \sum_{k \in \mathcal{K}_{\mathrm{e}}} \alpha_{\mathrm{e},k} \hat{R}_{\mathrm{e},k} \quad (10a)$$

$$\text{subject to } \|\boldsymbol{v}_k^{(t,w)}\| = 0, \ \forall t > T^{\max}/\tau - N + 1, t \in \mathcal{T}_0, \ \forall t > T^{\max}/\tau - N/2 + 1, t \in \mathcal{T}_1, \ \forall t > T^{\max}/\tau, t \in \mathcal{T}_2, \ k \in \mathcal{K}_{\mathrm{u}}, w \in \mathcal{W} \quad (10b)$$

$$\hat{R}_{\mathrm{u},k} \geq R^{\min}, \ k \in \mathcal{K}_{\mathrm{u}} \quad (10c)$$

$$\sum_{k \in \mathcal{K}} \sum_{t \in \mathcal{T}} \sum_{w \in \mathcal{W}} \|\boldsymbol{E}_m \boldsymbol{v}_k^{(t,w)}\|^2 \leq P^{\max}, m \in \mathcal{M} \quad (10d)$$

constraints (5b), (5c), and (9).

The expressions $\hat{R}_{\mathrm{e},k}$ and $\hat{R}_{\mathrm{u},k}$ are given as follows:

$$\hat{R}_{\mathrm{e},k} = \sum_{t \in \mathcal{T}} \sum_{w \in \mathcal{W}} \beta^{(t,w)} \tau^{(t,w)} \log_2 \left(1 + \gamma_k^{(t,w)}\right), k \in \mathcal{K}_{\mathrm{e}} \quad (11)$$

$$\hat{R}_{\mathrm{u},k} = \sum_{t \in \mathcal{T}} \sum_{w \in \mathcal{W}} \beta^{(t,w)} \tau^{(t,w)} \left(\log_2 \left(1 + \gamma_k^{(t,w)}\right) - \frac{Q^{-1}(\epsilon) \log_2 e}{\sqrt{L}} G_k^{(t,w)} \right), \ k \in \mathcal{K}_{\mathrm{u}}. \quad (12)$$

Furthermore, we introduce a set of auxiliary variables $c_k^{(t,w)}$ to URLLC user $k \in \mathcal{K}_{\mathrm{u}}$ using resource block (t, w), in order to bound the SINR:

$$0 \leq c_k^{(t,w)} \leq \gamma_k^{(t,w)}, \ k \in \mathcal{K}_{\mathrm{u}}, t \in \mathcal{T}, w \in \mathcal{W}. \quad (13)$$

Let vector $\boldsymbol{c}_k = (c_k^{(t,w)}, t \in \mathcal{T}, w \in \mathcal{W})$. We define $g(\boldsymbol{c}_k)$ as

$$g(\boldsymbol{c}_k) = \textstyle\sum_{t \in \mathcal{T}} \sum_{w \in \mathcal{W}} \beta^{(t,w)} \tau^{(t,w)} \left(\log_2 \left(1 + c_k^{(t,w)}\right) - \frac{Q^{-1}(\epsilon) \log_2 e}{\sqrt{L}} \sqrt{1 - (1 + c_k^{(t,w)})^{-2}} \right), \ k \in \mathcal{K}_{\mathrm{u}}. \quad (14)$$

Then, we can reformulate problem (10) as:

$$\underset{\substack{x^{(t,w)}, \boldsymbol{v}_k^{(t,w)}, c_k^{(t,w)} \\ k \in \mathcal{K}, t \in \mathcal{T}, w \in \mathcal{W}}}{\text{maximize}} \sum_{k \in \mathcal{K}_{\mathrm{e}}} \alpha_{\mathrm{e},k} \hat{R}_{\mathrm{e},k} \quad (15a)$$

$$\text{subject to } \ g(\boldsymbol{c}_k) \geq R^{\min}, \ k \in \mathcal{K}_{\mathrm{u}} \quad (15b)$$

constraints (5b), (5c), (9), (10b), (10d), (13).

Problem (15) is still difficult to solve due to the non-convexity of the SINR term in the objective function and constraint (13), the non-convexity of $g(\boldsymbol{c}_k)$ in constraint (15b), as well as the binary constraint (5b). We propose to tackle these issues by using fractional programming and SCA approaches. In particular, we first relax the binary constraint, and transform the SINR term into a convex term based on fractional programming. The problem can be reformulated as a difference of convex (DC) programming problem. Then, the nonconvex parts are approximated as convex terms using SCA technique.

To handle the binary constraint (5b), we relax it into $0 \leq x^{(t,w)} \leq 1$, $t \in \mathcal{T}, w \in \mathcal{W}$, and introduce a penalty term to the objective function. We have

$$\underset{\substack{x^{(t,w)}, \boldsymbol{v}_k^{(t,w)}, c_k^{(t,w)} \\ k \in \mathcal{K}, t \in \mathcal{T}, w \in \mathcal{W}}}{\text{maximize}} \sum_{k \in \mathcal{K}_{\mathrm{e}}} \alpha_{\mathrm{e},k} \hat{R}_{\mathrm{e},k} - \lambda \sum_{t \in \mathcal{T}} \sum_{w \in \mathcal{W}} \left(x^{(t,w)} - (x^{(t,w)})^2 \right) \quad (16a)$$

$$\text{subject to } \ 0 \leq x^{(t,w)} \leq 1, \ t \in \mathcal{T}, w \in \mathcal{W} \quad (16b)$$

constraints (5c), (9), (10b), (10d), (13), (15b),

where λ is a positive penalty coefficient. By using fractional programming [13] and introducing an auxiliary variable $\{z_k^{(t,w)}, k \in \mathcal{K}, t \in \mathcal{T}, w \in \mathcal{W}\}$, we can transform $\gamma_k^{(t,w)}$ into a convex expression with respect to the beamformer vector $\boldsymbol{v}_k^{(t,w)}$. Let $\boldsymbol{v} = (\boldsymbol{v}_k^{(t,w)}, k \in \mathcal{K}, t \in \mathcal{T}, w \in \mathcal{W})$ and $\boldsymbol{z} = (z_k^{(t,w)}, k \in \mathcal{K}, t \in \mathcal{T}, w \in \mathcal{W})$. The problem can be reformulated as follows:

$$\underset{\substack{x^{(t,w)}, \boldsymbol{v}_k^{(t,w)}, c_k^{(t,w)}, z_k^{(t,w)} \\ k \in \mathcal{K}, t \in \mathcal{T}, w \in \mathcal{W}}}{\text{maximize}} f(\boldsymbol{v}, \boldsymbol{z}) - \lambda \sum_{t \in \mathcal{T}} \sum_{w \in \mathcal{W}} \left(x^{(t,w)} - (x^{(t,w)})^2 \right) \quad (17a)$$

$$\text{subject to } \ 0 \leq c_k^{(t,w)} \leq \Gamma_k^{(t,w)}, k \in \mathcal{K}_{\mathrm{u}}, t \in \mathcal{T}, w \in \mathcal{W} \quad (17b)$$

constraints (5c), (9), (10b), (10d), (15b), (16b).

The objective function $f(\boldsymbol{v}, \boldsymbol{z})$ is defined as follows:

$$f(\boldsymbol{v}, \boldsymbol{z}) = \sum_{k \in \mathcal{K}_{\mathrm{e}}} \alpha_{\mathrm{e},k} \sum_{t \in \mathcal{T}} \sum_{w \in \mathcal{W}} \beta^{(t,w)} \tau^{(t,w)} \log_2 \left(1 + \Gamma_k^{(t,w)}\right), \quad (18)$$

and $\Gamma_k^{(t,w)}$ is given by

$$\Gamma_k^{(t,w)} = 2Re\left\{ (z_k^{(t,w)})^* (\boldsymbol{h}_k^{(t,w)})^H \boldsymbol{v}_k^{(t,w)} \right\} - |z_k^{(t,w)}|^2 \left(\sum_{l \in \mathcal{K} \setminus \{k\}} |(\boldsymbol{h}_k^{(t,w)})^H \boldsymbol{v}_l^{(t,w)}|^2 + (\sigma_k^{(t,w)})^2 \right). \quad (19)$$

The optimal solution of $z_k^{(t,w)}$ can be found in closed form [13], which leads to the equivalence between SINR and $\Gamma_k^{(t,w)}$:

$$z_k^{(t,w)} = \frac{(\boldsymbol{h}_k^{(t,w)})^H \boldsymbol{v}_k^{(t,w)}}{\sum_{l \in \mathcal{K} \setminus \{k\}} |(\boldsymbol{h}_k^{(t,w)})^H \boldsymbol{v}_l^{(t,w)}|^2 + (\sigma_k^{(t,w)})^2}. \quad (20)$$

Remark 1. To verify this transformation from problem (16) to (17), we can prove that the two problems achieve the same objective value by substituting (20) into (19), which leads to the equivalence of problems (16) and (17) [13].

The optimization variables in problem (17) can be updated iteratively. When $\{x^{(t,w)}, \boldsymbol{v}_k^{(t,w)}, c_k^{(t,w)}\}$ are fixed, the optimal $z_k^{(t,w)}$ is given by (20). When $z_k^{(t,w)}$ is fixed, the solution

to $\{x^{(t,w)}, \boldsymbol{v}_k^{(t,w)}, c_k^{(t,w)}\}$ can be determined by solving the following problem:

$$\underset{\substack{x^{(t,w)},\boldsymbol{v}_k^{(t,w)},c_k^{(t,w)}\\ k\in\mathcal{K},t\in\mathcal{T},w\in\mathcal{W}}}{\text{maximize}} f(\boldsymbol{v},\boldsymbol{z}) - \lambda\sum_{t\in\mathcal{T}}\sum_{w\in\mathcal{W}}\left(x^{(t,w)} - (x^{(t,w)})^2\right) \tag{21a}$$

$$\text{subject to constraints (5c), (9), (10b), (10d), (15b), (16b), and (17b),}$$

Problem (21) is a DC programming problem, where both the objective function and constraint (15b) are difference of convex functions. We can apply SCA to approximate the nonconvex terms by using Taylor series, and solve the reformulated convex problem through an iterative procedure.

Specifically, for the objective function (21a), we denote the nonconvex penalty term associated with variable $\boldsymbol{x} = (x^{(t,w)}, t\in\mathcal{T}, w\in\mathcal{W})$ as function $p(\boldsymbol{x})$:

$$p(\boldsymbol{x}) = \lambda\sum_{t\in\mathcal{T}}\sum_{w\in\mathcal{W}}\left(x^{(t,w)} - (x^{(t,w)})^2\right). \tag{22}$$

By employing SCA, we can determine the first-order approximation of function $p(\boldsymbol{x})$. In the j-th iteration, we have

$$\begin{aligned} p(\boldsymbol{x}) &\le p(\boldsymbol{x}^{(j)}) + \nabla p(\boldsymbol{x}^{(j)})^T(\boldsymbol{x}-\boldsymbol{x}^{(j)}) \\ &= \lambda\sum_{t\in\mathcal{T}}\sum_{w\in\mathcal{W}}\Big(x^{(t,w),(j)} - (x^{(t,w),(j)})^2 \\ &\quad + (1-2x^{(t,w),(j)})(x^{(t,w)} - x^{(t,w),(j)})\Big) \\ &\triangleq p^{(j)}(\boldsymbol{x}). \end{aligned} \tag{23}$$

For the expression $g(\boldsymbol{c}_k)$ in constraint (15b), we have

$$\begin{aligned} g(\boldsymbol{c}_k) &\le g(\boldsymbol{c}_k^{(j)}) + \nabla g(\boldsymbol{c}_k^{(j)})^T(\boldsymbol{c}_k - \boldsymbol{c}_k^{(j)}) \\ &= \sum_{t\in\mathcal{T}}\sum_{w\in\mathcal{W}}\beta^{(t,w)}\Bigg\{\log_2\left(1+c_k^{(t,w)}\right) \\ &\quad - \frac{Q^{-1}(\epsilon)\log_2 e}{\sqrt{N}}\Bigg[\sqrt{1-\left(1+c_k^{(t,w),(j)}\right)^{-2}} \\ &\quad + \frac{\left(1+c_k^{(t,w),(j)}\right)^{-3}}{\sqrt{1-\left(1+c_k^{(t,w),(j)}\right)^{-2}}}\left(c_k^{(t,w)} - c_k^{(t,w),(j)}\right)\Bigg]\Bigg\} \\ &\triangleq g^{(j)}(\boldsymbol{c}_k),\ k\in\mathcal{K}_{\text{u}}. \end{aligned} \tag{24}$$

By substituting (23) and (24) into problem (21), we can obtain the following convex problem to be solved in the j-th iteration:

$$\underset{\substack{x^{(t,w)},\boldsymbol{v}_k^{(t,w)},c_k^{(t,w)}\\ k\in\mathcal{K},t\in\mathcal{T},w\in\mathcal{W}}}{\text{maximize}} f(\boldsymbol{v},\boldsymbol{z}) - p^{(j)}(\boldsymbol{x}) \tag{25a}$$

$$\text{subject to } g^{(j)}(\boldsymbol{c}_k) \ge R^{\min}, k\in\mathcal{K}_{\text{u}} \tag{25b}$$

$$\text{constraints (5c), (9), (10b), (10d), (16b), (17b).}$$

Problem (25) is a convex optimization problem and can be solved by using solvers such as CVX. In order to obtain a feasible initialization which satifies the URLLC throughput requirement (25b), we propose to penalize the objective of problem (25) when constraint (25b) is violated. Problem (25) can be rewritten as

$$\underset{\substack{s_k,k\in\mathcal{K}_{\text{u}};x^{(t,w)},\boldsymbol{v}_k^{(t,w)}\\ c_k^{(t,w)},k\in\mathcal{K},t\in\mathcal{T},w\in\mathcal{W}}}{\text{maximize}} f(\boldsymbol{v},\boldsymbol{z}) - p^{(j)}(\boldsymbol{x}) - \xi\sum_{k\in\mathcal{K}_{\text{u}}} s_k \tag{26a}$$

$$\text{subject to } g^{(j)}(\boldsymbol{c}_k) + s_k \ge R^{\min}, k\in\mathcal{K}_{\text{u}} \tag{26b}$$

$$s_k \ge 0,\ k\in\mathcal{K}_{\text{u}} \tag{26c}$$

$$\text{constraints (5c), (9), (10b), (10d), (16b), (17b),}$$

where ξ is a penalty coefficient and $\{s_k, k\in\mathcal{K}_{\text{u}}\}$ are the slack variables. Problem (26) can be solved iteratively until convergence in polynomial time. Note that problem (26) is equivalent to problem (25) if feasible solutions exist and $s_k = 0$, for all $k\in\mathcal{K}_{\text{u}}$.

Algorithm 1 Proposed Algorithm for Resource Block Allocation and Beamforming Design

1: Initialize the variables $\boldsymbol{x}^{(0)}$, $\boldsymbol{v}^{(0)}$, $\boldsymbol{c}^{(0)}$, $\boldsymbol{z}^{(0)}$.
2: Initialize the thresholds δ_1 and δ_2.
3: Set $i := 0$.
4: **repeat**
5: $\quad i := i+1$.
6: $\quad$ Update $\boldsymbol{z}^{(i)}$ based on equation (20).
7: $\quad$ Set $j := 0$.
8: $\quad$ **repeat**
9: $\quad\quad j := j+1$.
10: $\quad\quad$ Solve problem (26) with fixed $\boldsymbol{z}$ to obtain the updated $\{\boldsymbol{x}^{(j)}, \boldsymbol{v}^{(j)}, \boldsymbol{c}^{(j)}\}$.
11: $\quad$ **until** $\|\boldsymbol{x}^{(j)}-\boldsymbol{x}^{(j-1)}\|+\|\boldsymbol{v}^{(j)}-\boldsymbol{v}^{(j-1)}\|+\|\boldsymbol{c}^{(j)}-\boldsymbol{c}^{(j-1)}\| \le \delta_1$.
12: **until** $\|\boldsymbol{z}^{(i)} - \boldsymbol{z}^{(i-1)}\| \le \delta_2$.
13: **Output:** $\boldsymbol{x}$, $\boldsymbol{v}$, $\boldsymbol{c}$, $\boldsymbol{z}$.

The proposed resource allocation algorithm is summarized in Algorithm 1. We denote $\boldsymbol{c} = (c_k^{(t,w)}, k\in\mathcal{K}, t\in\mathcal{T}, w\in\mathcal{W})$. The computational complexity of the proposed algorithm is dominated by the calculation of $\{\boldsymbol{x}^{(j)}, \boldsymbol{v}^{(j)}, \boldsymbol{c}^{(j)}\}$, i.e., Step 10 of Algorithm 1. Step 10 solves a convex optimization problem with computational complexity in the order of $O\left(I_1 I_2(|\mathcal{T}||\mathcal{W}|(K_{\text{e}}+K_{\text{u}})M)^4 N_{\text{t}}^3)\right)$ [12], where I_1 and I_2 are the number of iterations for the outer and inner loops in Algorithm 1, respectively.

IV. Performance Evaluation

In this section, we present simulation results to evaluate the performance of the proposed algorithm. We consider there are $M = 3$ BSs, located in an area with radius 0.5 km. The BSs are located in the middle of the coverage area. The distance between a pair of BSs is the same. Users are randomly distributed within the coverage area. Each BS is equipped with $N_{\text{t}} = 8$ antennas. The path-loss is set as $L(d_{k,m}) = 128.1 + 37.6\log(d_{k,m})$, where $d_{k,m}$ is the distance between user k and BS m. The small-scale fading between each BS and user follows an independent and identically Rayleigh distribution (i.e., $\mathcal{CN}(0,1)$). We set the maximum BS transmit power $P^{\max}$ and the noise power density as 45 dBm and -174 dBm/Hz, respectively. Similar to [6] and [10], we consider the radio resource with total bandwidth $W = 2160$ kHz and time duration $T = 1.5$ ms, where the size of the

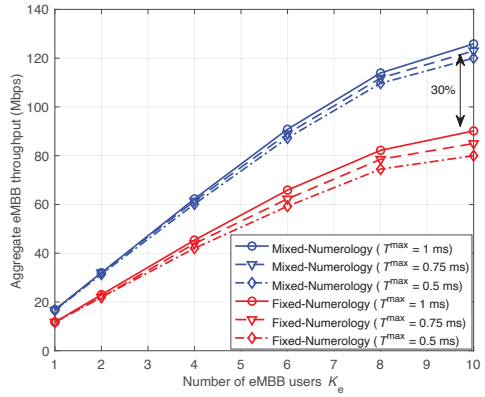

Fig. 3. Aggregate throughput of eMBB users versus the number of eMBB users K_e ($K_u = K_e$, $P^{max} = 45$ dBm, and $R^{min} = 500$ kbps).

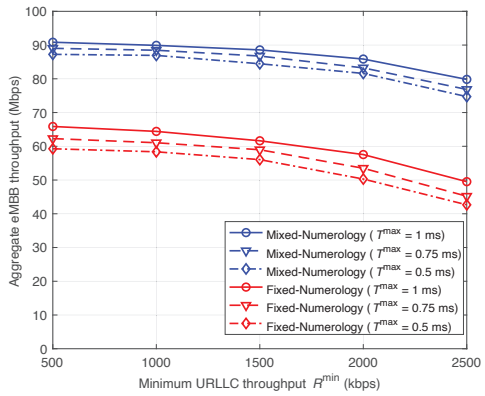

Fig. 4. Aggregate throughput of eMBB users versus the minimum throughput requirement of URLLC users R^{min} ($K_u = K_e = 6$, $P^{max} = 45$ dBm).

resource grid is 12×12. We set the weights $\alpha_{e,k}$ and $\alpha_{u,k}$ to be equal to 1. The penalty weights are set as follows: $\lambda = 5000$ and $\xi = 1000$.

In Fig. 3, we show the aggregate throughput of eMBB users versus the number of eMBB users. We set the number of URLLC users to be equal to the number of eMBB users, i.e., $K_u = K_e$. We set the minimum throughput requirement of URLLC users R^{min} to be 500 kbps. We consider the latency requirements to be $T^{max} = \{0.5, 0.75, 1\}$ ms. We include the fixed-numerology with SCS of 15 kHz used in LTE system for performance comparison. Results in Fig. 3 show that the proposed mixed-numerology approach has more performance advantages when there are more users in the system or the URLLC latency requirement becomes more stringent. When $K_e = K_u = 10$, our proposed algorithm can improve the aggregate eMBB throughput by 30% compared with the fixed-numerology approach.

Fig. 4 shows the aggregate throughput of eMBB users versus the minimum throughput requirement of URLLC users R^{min}. The number of eMBB and URLLC users are set as $K_e = K_u = 6$. When R^{min} increases, the aggregate eMBB throughput is reduced. Our proposed algorithm still outperforms the fixed-numerology approach. This performance advantage is brought by the flexibility of mixed-numerology. The proposed algorithm with mixed-numerology has more choices in selecting the resource blocks to improve the eMBB throughput while guaranteeing the QoS requirements of URLLC users.

V. Conclusion

In this paper, we investigated the joint beamforming and resource block allocation with mixed-numerology for eMBB and URLLC services. We formulated a mixed-integer nonlinear problem which aims at maximizing the aggregate throughput of eMBB users while guaranteeing the latency, reliability, and throughput requirements of URLLC users. The formulated problem is nonconvex due to the inter-user interference and the short packet transmission of URLLC users. To solve this nonconvex problem, we proposed a suboptimal solution with polynomial complexity by using fractional programming and SCA techniques. We compared the simulation results of the proposed algorithm with the fixed-numerology approach. Results showed that the proposed algorithm can improve the aggregate eMBB throughput by 30% when compared with the fixed-numerology based approach. For future work, we will consider multi-antenna at the user side and include mMTC use case in the problem formulation.

References

[1] M. Shafi *et al.*, "5G: A tutorial overview of standards, trials, challenges, deployment, and practice," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 6, pp. 1201−1221, Jun. 2017.

[2] J. Tang, B. Shim, and T. Q. S. Quek, "Service multiplexing and revenue maximization in sliced C-RAN incorporated with URLLC and multicast eMBB," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 4, pp. 881−895, Apr. 2019.

[3] M. Setayesh, S. Bahrami, and V. W. S. Wong, "Joint PRB and power allocation for slicing eMBB and URLLC services in 5G C-RAN," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Taipei, Taiwan, Dec. 2020.

[4] A. Anand, G. D. Veciana, and S. Shakkottai, "Joint scheduling of URLLC and eMBB traffic in 5G wireless networks," *IEEE/ACM Trans. Netw.*, vol. 28, no. 2, pp. 477−490, Apr. 2020.

[5] *Study on New Radio (NR) Access Technology (Release 16)*, document 3GPP TR 38.912, Jul. 2020.

[6] L. You, Q. Liao, N. Pappas, and D. Yuan, "Resource optimization with flexible numerology and frame structure for heterogeneous services," *IEEE Commun. Lett.*, vol. 22, no. 12, pp. 2579−2582, Dec. 2018.

[7] 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.*, vol. 24, no. 2, pp. 410−413, Feb. 2020.

[8] C. Tang, X. Chen, Y. Chen, and Z. Li, "Dynamic resource optimization based on flexible numerology and Markov decision process for heterogeneous services," in *Proc. IEEE Int'l Conf. on Parallel and Distrib. (ICPADS)*, Tianjin, China, Dec. 2019.

[9] P. K. Korrai, E. Lagunas, A. Bandi, S. K. Sharma, and S. Chatzinotas, "Joint power and resource block allocation for mixed-numerology-based 5G downlink under imperfect CSI," *IEEE Open J. Commun. Soc.*, vol. 1, pp. 1583−1601, Oct. 2020.

[10] W. Sui, X. Chen, S. Zhang, Z. Jiang, and S. Xu, "Energy-efficient resource allocation with flexible frame structure for hybrid eMBB and URLLC services," *IEEE Trans. Green Commun. and Netw.*, vol. 15, no. 1, pp. 72−83, Mar. 2021.

[11] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307−2359, May 2010.

[12] W. R. Ghanem, V. Jamali, Y. Sun, and R. Schober, "Resource allocation for multi-user downlink MISO OFMDA-URLLC systems," *IEEE Trans. Commun.*, vol. 68, no. 11, pp. 7184−7200, Aug. 2020.

[13] K. Shen and W. Yu, "Fractional programming for communication systems−Part I: Power control and beamforming," *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2616−2630, Mar. 2018.