Joint PRB and Power Allocation for Slicing eMBB and URLLC Services in 5G C-RAN

Mehdi Setayesh, Shahab Bahrami, and Vincent W.S. Wong
Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, Canada
email: {setayeshm, bahramis, vincentw}@ece.ubc.ca

***Abstract*—Efficient allocation of resources (i.e., physical resource blocks (PRBs), transmit power) for remote radio heads (RRHs) in the fifth generation (5G) cloud radio access network (C-RAN) is crucial for the mobile network operators (MNOs) to support different use cases with diverse quality of service (QoS) requirements. In this paper, we study the resource allocation of enhanced mobile broadband (eMBB) and ultra-reliable low-latency communications (URLLC) network slices in a 5G C-RAN. We formulate the resource allocation problem as a mixed-integer nonlinear program. We address the isolation between eMBB and URLLC network slices and the uncertainty in the traffic load by using the chance constraint. We consider short packet transmission to enable URLLC data transmission with low latency and high reliability. We propose an algorithm based on penalized successive convex approximation to determine a suboptimal solution of the formulated problem. The proposed algorithm has a polynomial time complexity. Simulation results show that the proposed algorithm on average achieves 30% higher throughput when compared with a baseline scheme that only optimizes the transmit power of users.**

I. Introduction

The fifth generation (5G) wireless systems aim to support different use cases such as enhanced mobile broadband (eMBB), massive machine-type communication (mMTC), and ultra-reliable low-latency communications (URLLC) with various quality of service (QoS) requirements in terms of data rate, reliability, and latency. Mobile network operators (MNOs) can use network slicing in cloud radio access networks (C-RANs) to virtualize the network resources (e.g., physical resource blocks (PRBs), transmit power) in the shared physical network. Given the limited resources in the remote radio heads (RRHs), an efficient resource allocation among network slices is crucial to jointly maximize the aggregate system throughput subject to the users' QoS requirements [1].

There are major challenges in allocating RRH resources among different network slices. First, the MNO has uncertainty in the traffic load and the arrival rate of the users' traffic. Second, the MNO needs to consider the trade-off between maximizing the utility (e.g., aggregate throughput) and allocating sufficient RRH resources per slice to guarantee isolation between slices. Moreover, the MNO should take into account the interference between the RRHs when allocating the same PRBs to the users within a given coverage area.

Recently, the coexistence of eMBB and URLLC traffic in a shared C-RAN infrastructure has received considerable attention. To meet the QoS requirements of such traffic, two types of scheduling algorithms have been proposed in the literature: punctured scheduling and orthogonal scheduling [2]. In the punctured scheduling approach, the eMBB traffic is suspended when the URLLC packets are being sent. Thus, the radio resources for the eMBB traffic are reallocated to the URLLC users to meet their stringent latency requirements. Pandey *et al.* [3] applied the punctured scheduling approach and proposed a dynamic resource allocation algorithm to maximize the aggregate data rate of the eMBB users, while maintaining the low latency for the URLLC users. Alsenwi *et al.* [4] proposed a risk-sensitive punctured scheduling approach by reallocating the radio resources for the eMBB users with high data rate to the URLLC users.

In the punctured scheduling approach, the decoding performance for punctured eMBB packets degrades. To address this shortcoming, radio resources are reserved for eMBB and URLLC services in the orthogonal scheduling approach. Ma *et al.* [5] considered network slicing in a downlink orthogonal frequency division multiple access (OFDMA) based network to maximize the spectral efficiency. Hua *et al.* [6] applied deep reinforcement learning to design an online resource allocation algorithm. Lee *et al.* [7] proposed a bi-level network slicing framework, where an admission control is used in the first level and resource allocation among the admitted users is performed in the second level. Although the proposed approaches in [5]–[7] can be applied to allocate RRH resources among different slices to meet the required QoS of the users, they do not consider isolation between network slices and short data packet transmission for URLLC traffic. The proposed algorithms in [5], [6] are designed for one cell, which cannot be directly used in a network with multiple RRHs. Moreover, the work in [7] considered deterministic traffic load and interference for RRHs, which may not be applicable to systems with stochastic traffic.

In this paper, we study the radio resource allocation problem in an OFDMA-based C-RAN with multiple RRHs. We develop a radio resource allocation algorithm for eMBB and URLLC traffic by taking into account the uncertainty in the traffic load, isolation between network slices, and the interference between RRHs. The main contributions of this paper are as follows:

- *Traffic Load Uncertainty:* We capture the uncertainty in the arrival rate and packet size of the users' traffic load in the eMBB and URLLC slices as a chance constraint. The tunable parameter in the chance constraint enables us to allocate sufficient resources to meet the aggregate traffic

load and guarantee isolation between network slices.

- *URLLC QoS Requirements:* We model the traffic of the URLLC users with finite blocklength data rate. It enables us to consider short packet transmission to address the low latency requirement of the URLLC users.
- *Algorithm Design:* We take into account the interference of the neighboring RRHs and formulate the joint PRBs and transmit power allocation among network slices as a mixed-integer nonlinear program. We apply the penalized successive convex approximation technique and develop an algorithm with polynomial time complexity that converges to a suboptimal solution of the formulated resource allocation problem.
- *Performance Evaluation:* We perform simulations with multiple RRHs serving eMBB and URLLC slices. Results show that the proposed algorithm achieves 30% higher aggregate throughput when compared with a baseline scheme, where the PRBs are allocated randomly and the MNO optimizes the transmit power. When only eMBB traffic is considered, our proposed algorithm also provides a higher aggregate throughput than another recently proposed algorithm [7].

The remainder of this paper is organized as follows. The system model is described in Section II. The details of our proposed resource allocation algorithm is presented in Section III. In Section IV, we evaluate the performance of the proposed algorithm. Section V concludes the paper.

II. System Model

Consider a C-RAN shown in Fig. 1 that consists of H RRHs, one baseband unit (BBU), and U users. Let $\mathcal{H} = \{1, \ldots, H\}$ denote the set of RRHs and $\mathcal{U} = \{1, \ldots, U\}$ denote the set of users. We assume that user association is given. Hence, we denote the set of users associated with RRH $h \in \mathcal{H}$ by $\mathcal{U}_h \subseteq \mathcal{U}$. The RRHs are connected to the BBU via the fronthaul links. Similar to [7], we assume shared spectrum mode where all RRHs can access the same set $\mathcal{K} = \{1, \ldots, K\}$ of K PRBs. Since a large portion of the network traffic is constituted by the downlink traffic [8], we focus on resource allocation in downlink direction in this paper. Let $\mathcal{S} = \{1, \ldots, S\}$ denote the set of network slices. We divide the set of slices into the set $\mathcal{S}^{\text{eMBB}} \subseteq \mathcal{S}$ of eMBB slice type and the set $\mathcal{S}^{\text{URLLC}} \subseteq \mathcal{S}$ of URLLC slice type. Each user belongs to one network slice based on its required services. Let $\mathcal{U}_s \subseteq \mathcal{U}$ denote the set of users associated with slice $s \in \mathcal{S}$. Each network slice s guarantees the QoS of the admitted users in set $\mathcal{U}_s$. We define set $\mathcal{U}_{s,h} = \mathcal{U}_s \cap \mathcal{U}_h$ to be the set of $U_{s,h}$ users in slice s that are served by RRH h.

A. QoS Constraints of the Users

The RRH resources (i.e., PRBs, transmit power) are shared between the users within the coverage area. We use the binary decision variable $m_{u,k}$ to indicate whether PRB $k \in \mathcal{K}$ is allocated to user $u \in \mathcal{U}$ ($m_{u,k} = 1$) or not ($m_{u,k} = 0$). We use the continuous decision variable $p_{u,k}$ to denote the transmit power from RRH $h \in \mathcal{H}$ to its serving user $u \in \mathcal{U}_h$

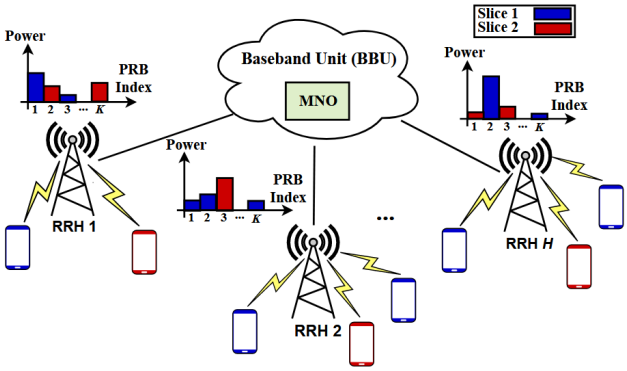

Fig. 1. Illustration of a C-RAN. The MNO allocates the RRHs' resources (i.e., PRBs, transmit power) to the users in each slice.

using PRB k. We denote the maximum transmit power of RRH $h \in \mathcal{H}$ by $P_h^{\max}$. We have

$$\sum_{u \in \mathcal{U}_h} \sum_{k \in \mathcal{K}} m_{u,k}\, p_{u,k} \leq P_h^{\max}, \qquad h \in \mathcal{H}, \tag{1a}$$

$$p_{u,k} \geq 0, \qquad u \in \mathcal{U}_h, \quad h \in \mathcal{H}, \quad k \in \mathcal{K}, \tag{1b}$$

$$m_{u,k} \in \{0, 1\}, \qquad u \in \mathcal{U}_h, \quad h \in \mathcal{H}, \quad k \in \mathcal{K}. \tag{1c}$$

Each PRB $k \in \mathcal{K}$ can be allocated to at most one user within the coverage area of RRH h. We have

$$\sum_{u \in \mathcal{U}_h} m_{u,k} \leq 1, \qquad h \in \mathcal{H}, \quad k \in \mathcal{K}. \tag{2}$$

The PRB $k \in \mathcal{K}$ can be allocated to the users in different RRHs. Hence, a user in RRH h using PRB k can experience interference from other RRHs in set $\mathcal{H} \backslash \{h\}$. We consider an upper bound $I_s^{\max}$ for the acceptable interference experienced by the users in slice $s \in \mathcal{S}$. The value of parameter $I_s^{\max}$ is determined according to the QoS requirements of the users in slice s. Let $g_{u,h,k} \in \mathbb{C}$ denote channel gain between user u and RRH h on PRB k. For user $u \in \mathcal{U}_{s,h}$ in network slice $s \in \mathcal{S}$ served by RRH $h \in \mathcal{H}$ using PRB $k \in \mathcal{K}$, we have

$$m_{u,k} \sum_{h' \in \mathcal{H} \backslash \{h\}} \sum_{u' \in \mathcal{U}_{h'}} |g_{u,h',k}|^2\, m_{u',k}\, p_{u',k} \leq I_s^{\max}. \tag{3}$$

The signal-to-noise plus interference ratio (SINR) experienced by user $u \in \mathcal{U}_h$ served by RRH $h \in \mathcal{H}$ using PRB $k \in \mathcal{K}$ is as follows:

$$\Gamma_{u,k} = \frac{|g_{u,h,k}|^2\, p_{u,k}}{\sum_{h' \in \mathcal{H} \backslash \{h\}} \sum_{u' \in \mathcal{U}_{h'}} |g_{u,h',k}|^2\, m_{u',k}\, p_{u',k} + \sigma^2}, \tag{4}$$

where σ^2 is the variance of the additive white Gaussian noise.

Next we obtain the users' data rate in eMBB and URLLC slices. Let $C(\Gamma_{u,k}) = B \log_2(1 + \Gamma_{u,k})$ denote the Shannon capacity of the communication channel between user $u \in \mathcal{U}_h$ and RRH $h \in \mathcal{H}$ on PRB $k \in \mathcal{K}$, where B is the bandwidth of PRB k. The data rate for user $u \in \mathcal{U}_{s,h}$ in eMBB slice $s \in \mathcal{S}^{\text{eMBB}}$ served by RRH $h \in \mathcal{H}$ using PRB $k \in \mathcal{K}$ is as follows:

$$r_{u,k} = C(\Gamma_{u,k}). \tag{5}$$

Due to the finite blocklength N_u^{b} in URLLC traffic, the Shannon capacity cannot be used to obtain the data rate for the URLLC users. We consider short packet transmission for URLLC users. Hence, the data rate for user $u \in \mathcal{U}_{s,h}$ in slice $s \in \mathcal{S}^{\mathrm{URLLC}}$ served by RRH $h \in \mathcal{H}$ using PRB $k \in \mathcal{K}$ can be approximated as follows [9]:

$$r_{u,k} = C(\Gamma_{u,k}) - D(\Gamma_{u,k}), \tag{6a}$$

where

$$D(\Gamma_{u,k}) = B \log_2 e \, Q^{-1}(\epsilon_u^{\mathrm{b}}) \sqrt{\frac{V_{u,k}}{N_u^{\mathrm{b}}}}. \tag{6b}$$

In (6b), $V_{u,k} = \frac{\Gamma_{u,k}}{1+\Gamma_{u,k}}$ represents the channel dispersion, $Q^{-1}(\cdot)$ is the inverse of the Gaussian Q-function, and ϵ_u^{b} denotes the transmission error probability.

We characterize the traffic load of user $u \in \mathcal{U}$ as a random variable $L_u = X_u \Lambda_u$, where X_u and Λ_u represent the packet size (in bits) and the packet arrival rate (in packets per second), respectively. To guarantee the QoS of user u in slice s, the probability that the traffic load exceeds the allocated total data rate for user u should be less than or equal to the maximum tolerable probability ϵ_s of failure on supporting the traffic load, where $0 < \epsilon_s < 1$. We have

$$\Pr\Big(L_u > \sum_{k \in \mathcal{K}} m_{u,k} \, r_{u,k}\Big) \leq \epsilon_s, \quad u \in \mathcal{U}_s, \ s \in \mathcal{S}. \tag{7}$$

As recommended by the Third Generation Partnership Project (3GPP) specification, a range of realistic traffic types for eMBB and URLLC slices can be captured by tuning the parameters of file transfer protocol (FTP)-based models [10]. In this paper, we model the traffic patterns from eMBB and URLLC users as the FTP model 3 in [11]. In particular, for each user u, the packet size X_u is a constant and the packet arrival rate Λ_u follows the Poisson distribution with parameter λ_u. Let $F_{\Lambda_u}(\cdot)$ denote the cumulative distribution function (CDF) of the packet arrival rate for user u. When Λ_u follows the Poisson distribution, we can express $F_{\Lambda_u}(\cdot)$ in terms of incomplete gamma functions [12]. By performing some algebraic manipulations, we can express inequality (7) as:

$$\sum_{k \in \mathcal{K}} m_{u,k} \, r_{u,k} \geq X_u F_{\Lambda_u}^{-1}(1 - \epsilon_s), \quad u \in \mathcal{U}_s, \ s \in \mathcal{S}, \tag{8}$$

where $F_{\Lambda_u}^{-1}(\cdot)$ is the inverse of CDF $F_{\Lambda_u}(\cdot)$.

B. Network Slices Isolation and Problem Formulation

To provide isolation between network slices, the MNO needs to guarantee the minimum required RRH resources (i.e., PRBs, transmit power) for all the users $u \in \mathcal{U}_{s,h}$ in slice $s \in \mathcal{S}$ served by RRH $h \in \mathcal{H}$ based on their aggregate traffic load. Let δ_s denote the maximum tolerable probability of failure on supporting the aggregate traffic load of all the users in slice $s \in \mathcal{S}$ and RRH $h \in \mathcal{H}$. We have

$$\Pr\Big(\sum_{u \in \mathcal{U}_{s,h}} L_u > \sum_{u \in \mathcal{U}_{s,h}} \sum_{k \in \mathcal{K}} m_{u,k} \, r_{u,k}\Big) \leq \delta_s. \tag{9}$$

By tuning parameter δ_s, MNO can guarantee more resources for the users in slice s. In order to simplify further calculations, we use $\overline{X}_{s,h}$ to denote the average packet size of all the users $u \in \mathcal{U}_{s,h}$ in slice $s \in \mathcal{S}$ served by RRH $h \in \mathcal{H}$. Let $\Lambda_{s,h}$ denote the data arrival rate for the aggregate load of the users in slice s served by RRH h. Since $\Lambda_{s,h}$ is the sum of independent Poisson random variables, it is also a Poisson random variable with parameter $\sum_{u \in \mathcal{U}_{s,h}} \lambda_u$. By performing some algebraic manipulations, inequality (9) for slice $s \in \mathcal{S}$ and RRH $h \in \mathcal{H}$ can be expressed as follows:

$$\sum_{u \in \mathcal{U}_{s,h}} \sum_{k \in \mathcal{K}} m_{u,k} \, r_{u,k} \geq \overline{X}_{s,h} F_{\Lambda_{s,h}}^{-1}(1 - \delta_s). \tag{10}$$

We assign a priority factor α_s to each slice $s \in \mathcal{S}$, where $0 \leq \alpha_s \leq 1$ and $\sum_{s \in \mathcal{S}} \alpha_s = 1$. The value of α_s is determined based on the service level agreement between the slice owner and the MNO [13]. Furthermore, we introduce another priority factor β_u for user u in each slice $s \in \mathcal{S}$, where $0 \leq \beta_u \leq 1$ and $\sum_{u \in \mathcal{U}_s} \beta_u = 1$. The C-RAN resource allocation problem can be formulated as follows:

$$\underset{\substack{m_{u,k},\, p_{u,k},\\ u \in \mathcal{U},\ k \in \mathcal{K}}}{\text{maximize}} \quad \sum_{s \in \mathcal{S}} \sum_{u \in \mathcal{U}_s} \sum_{k \in \mathcal{K}} \alpha_s \beta_u \, m_{u,k} \, r_{u,k} \tag{11}$$

subject to constraints (1a)$-$(3), (8), and (10).

Problem (11) is a mixed-integer nonlinear program, which is NP-hard and difficult to solve. In the next section, we propose an algorithm with a polynomial time complexity to determine a suboptimal solution to problem (11).

III. Proposed Algorithm

In this section, we apply a penalized successive convex approximation approach to determine a suboptimal solution of problem (11). For user $u \in \mathcal{U}$ and PRB $k \in \mathcal{K}$, we define an auxiliary variable $\widetilde{p}_{u,k}$ as follows:

$$\widetilde{p}_{u,k} = m_{u,k} \, p_{u,k}, \quad u \in \mathcal{U}, \ \ k \in \mathcal{K}. \tag{12}$$

We can express constraints (1a) and (3) in terms of the auxiliary variables $\widetilde{p}_{u,k}$, $u \in \mathcal{U}$, $k \in \mathcal{K}$. Considering this, we can rewrite (3) in the following equivalent form:

$$\sum_{h' \in \mathcal{H} \setminus \{h\}} \sum_{u' \in \mathcal{U}_{h'}} |g_{u,h',k}|^2 \, \widetilde{p}_{u',k} \leq m_{u,k} I_s^{\max} + (1 - m_{u,k}) M, \tag{13}$$

where M is a sufficiently large number. Furthermore, we can rewrite the objective function as well as constraints (8) and (10) in problem (11) using the following properties:

$$m_{u,k} r_{u,k} = C(m_{u,k}\Gamma_{u,k}), \ u \in \mathcal{U}_s, \ s \in \mathcal{S}^{\mathrm{eMBB}}, \ k \in \mathcal{K}, \tag{14a}$$

$$m_{u,k} r_{u,k} = C(m_{u,k}\Gamma_{u,k}) - D(m_{u,k}\Gamma_{u,k}), \quad u \in \mathcal{U}_s, \ s \in \mathcal{S}^{\mathrm{URLLC}}, \ k \in \mathcal{K}. \tag{14b}$$

We introduce the following inequalities in the constraints set to decompose the product terms in (12) using the big-M approach [14]:

$$\widetilde{p}_{u,k} \leq M m_{u,k}, \quad u \in \mathcal{U}, \ \ k \in \mathcal{K}, \tag{15a}$$

$$\widetilde{p}_{u,k} \leq p_{u,k}, \quad u \in \mathcal{U}, \ \ k \in \mathcal{K}, \tag{15b}$$

$$\widetilde{p}_{u,k} \geq p_{u,k} - (1 - m_{u,k})M, \qquad u \in \mathcal{U},\ \ k \in \mathcal{K}, \tag{15c}$$
$$\widetilde{p}_{u,k} \geq 0, \qquad u \in \mathcal{U},\ \ k \in \mathcal{K}. \tag{15d}$$

Next, we relax the nonconvex constraint (1c) in the form of difference of two convex functions. We have

$$\sum_{u\in\mathcal{U}}\sum_{k\in\mathcal{K}} m_{u,k} - \sum_{u\in\mathcal{U}}\sum_{k\in\mathcal{K}} (m_{u,k})^2 \leq 0, \tag{16a}$$
$$0 \leq m_{u,k} \leq 1, \qquad u \in \mathcal{U},\ k \in \mathcal{K}. \tag{16b}$$

To address the nonlinear structure of $\Gamma_{u,k}$ and constraints (8) and (10) in problem (11), we introduce auxiliary variables $Z_{u,k}$, $u \in \mathcal{U}_h$, $h \in \mathcal{H}$, $k \in \mathcal{K}$, such that:

$$0 \leq Z_{u,k} \leq \frac{|g_{u,h,k}|^2\, \widetilde{p}_{u,k}}{\sum_{h'\in\mathcal{H}\setminus\{h\}} \sum_{u'\in\mathcal{U}_{h'}} |g_{u,h',k}|^2\, \widetilde{p}_{u',k} + \sigma^2}. \tag{17}$$

We include inequality (17) in the constraints set and express functions $C(\cdot)$ and $D(\cdot)$ in terms of variables $Z_{u,k}$, $u \in \mathcal{U}_h$, $h \in \mathcal{H}$, $k \in \mathcal{K}$. We can show that $C(Z_{u,k})$ and $D(Z_{u,k})$ are monotonic concave functions when ϵ_u^b is less than 0.5 for all the URLLC users. Therefore, constraint (17) holds with equality in the optimal solution, i.e., $Z_{u,k} = m_{u,k}\Gamma_{u,k}$. Constraint (17) is nonconvex. We rewrite (17) as follows:

$$\Big(\sum_{h'\in\mathcal{H}\setminus\{h\}} \sum_{u'\in\mathcal{U}_{h'}} |g_{u,h',k}|^2 \widetilde{p}_{u',k} + \sigma^2\Big) Z_{u,k} \leq |g_{u,h,k}|^2\, \widetilde{p}_{u,k}, \qquad u \in \mathcal{U}_h,\ h \in \mathcal{H},\ k \in \mathcal{K}. \tag{18}$$

We use equality $f_1(x)f_2(x) = \frac{1}{2}((f_1(x)+f_2(x))^2 - f_1(x)^2 - f_2(x)^2)$ to express the left-hand side of (18) as a difference of two convex functions [15]. The objective function and constraints of problem (11) are transformed into a convex function or a difference of two convex functions. Finally, we define the following functions:

$$\Psi(m_{u,k}) = (m_{u,k})^2, \tag{19a}$$
$$\Theta_1(Z_{u,k}, \widetilde{\boldsymbol{p}}) = \frac{1}{2}\Big(Z_{u,k} + \sum_{h'\in\mathcal{H}\setminus\{h\}} \sum_{u'\in\mathcal{U}_{h'}} |g_{u,h',k}|^2 \widetilde{p}_{u',k}\Big)^2, \tag{19b}$$
$$\Theta_2(Z_{u,k}, \widetilde{\boldsymbol{p}}) = \frac{1}{2}\Big((Z_{u,k})^2 + \big(\sum_{h'\in\mathcal{H}\setminus\{h\}} \sum_{u'\in\mathcal{U}_{h'}} |g_{u,h',k}|^2 \widetilde{p}_{u',k}\big)^2\Big), \tag{19c}$$

where vector $\widetilde{\boldsymbol{p}} = (\widetilde{p}_{u,k},\ u \in \mathcal{U}_h,\ h \in \mathcal{H},\ k \in \mathcal{K})$.

Algorithm 1 describes our proposed resource allocation algorithm to solve problem (11) for the eMBB and URLLC slices. Algorithm 1 is based on the penalized successive convex approximation [16]. We accept violations from constraints (8), (10), (16a), and (18) and use a penalizing weight coefficient ζ, which increases gradually by a factor of η. We denote the maximum penalizing weight coefficient by $\zeta^{\max}$.

Let i denote the iteration index. In Algorithm 1, Lines 1 to 3 describe the initialization for the maximum number of iterations $i^{\max}$, the algorithm's hyperparameters $\zeta^{(1)}$, $\zeta^{\max}$, and η, as well as the decision variables $m_{u,k}^{(1)}$, $\widetilde{p}_{u,k}^{(1)}$, and $Z_{u,k}^{(1)}$ for user $u \in \mathcal{U}$ and PRB $k \in \mathcal{K}$ in iteration $i = 1$.

Algorithm 1 Proposed Algorithm Based on Penalized Successive Convex Approximation

1: Set $i := 1$ and initialize the maximum number of iterations $i^{\max}$.
2: MNO initializes hyperparameters $\zeta^{(1)}$ and $\zeta^{\max}$, and η.
3: MNO randomly initializes variables $m_{u,k}^{(1)}$, $\widetilde{p}_{u,k}^{(1)}$, and $Z_{u,k}^{(1)}$, for user $u \in \mathcal{U}$ and PRB $k \in \mathcal{K}$.
4: **Repeat**
5: MNO determines the optimal solution $m_{u,k}^*$, $\widetilde{p}_{u,k}^*$, and $Z_{u,k}^*$ of problem (21).
6: MNO updates variables $m_{u,k}^{(i+1)} := m_{u,k}^*$, $\widetilde{p}_{u,k}^{(i+1)} := \widetilde{p}_{u,k}^*$, and $Z_{u,k}^{(i+1)} := Z_{u,k}^*$ for user $u \in \mathcal{U}$ and PRB $k \in \mathcal{K}$.
7: $\zeta^{(i+1)} := \min\{\eta\,\zeta^{(i)}, \zeta^{\max}\}$.
8: Set $i := i + 1$.
9: **Until** $i = i^{\max}$ or $m_{u,k}$, $\widetilde{p}_{u,k}$, and $Z_{u,k}$ converge.
10: MNO sets $m_{u,k}^{\text{opt}} := m_{u,k}^{(i)}$, $\widetilde{p}_{u,k}^{\text{opt}} := \widetilde{p}_{u,k}^{(i)}$, $Z_{u,k}^{\text{opt}} := Z_{u,k}^{(i)}$ for user $u \in \mathcal{U}$ and PRB $k \in \mathcal{K}$.

In the loop within Lines 4 to 9, the MNO solves a convex optimization problem iteratively. In Line 5, the MNO uses the optimal solution $m_{u,k}^*$, $\widetilde{p}_{u,k}^*$, and $Z_{u,k}^*$ in the previous iteration to obtain the first-order approximations $\widehat{D}(Z_{u,k})$, $\widehat{\Psi}(m_{u,k})$, and $\widehat{\Theta}_2(Z_{u,k}, \widetilde{\boldsymbol{p}})$ for functions $D(Z_{u,k})$, $\Psi(m_{u,k})$, and $\Theta_2(Z_{u,k}, \widetilde{\boldsymbol{p}})$, respectively. The violations from constraints (8), (10), (16a), and (18) are penalized by the weighting coefficient $\zeta^{(i)}$. The objective function is as follows:

$$f^{\text{obj}} = \sum_{k\in\mathcal{K}}\Big(\sum_{s\in\mathcal{S}}\sum_{u\in\mathcal{U}_s} \alpha_s\beta_u C(Z_{u,k}) - \sum_{s\in\mathcal{S}^{\text{URLLC}}}\sum_{u\in\mathcal{U}_s} \alpha_s\beta_u \widehat{D}(Z_{u,k})\Big) - \zeta^{(i)}\Big(\tau + \sum_{u\in\mathcal{U}}\sum_{k\in\mathcal{K}} \tau_{u,k} + \sum_{u\in\mathcal{U}} \tau_u + \sum_{s\in\mathcal{S}}\sum_{h\in\mathcal{H}} \tau_{s,h}\Big). \tag{20}$$

The MNO solves the following optimization problem:

$$\underset{\substack{m_{u,k}, p_{u,k}, \widetilde{p}_{u,k},\\ Z_{u,k}, \tau, \tau_{u,k}, \tau_u, \tau_{s,h},\\ u\in\mathcal{U},\, k\in\mathcal{K},\, s\in\mathcal{S},\, h\in\mathcal{H}}}{\text{maximize}} \quad f^{\text{obj}} \tag{21}$$

subject to constraints (1b), (2), (13), (15a)$-$(15d), and (16b),

$$\sum_{u\in\mathcal{U}}\sum_{k\in\mathcal{K}} m_{u,k} - \sum_{u\in\mathcal{U}}\sum_{k\in\mathcal{K}} \widehat{\Psi}(m_{u,k}) - \tau \leq 0,$$
$$\sum_{k\in\mathcal{K}} C(Z_{u,k}) + \tau_u \geq X_u F_{\Lambda_u}^{-1}(1-\epsilon_s),\ u \in \mathcal{U}_s,\ s \in \mathcal{S}^{\text{eMBB}},$$
$$\sum_{k\in\mathcal{K}} C(Z_{u,k}) - \sum_{k\in\mathcal{K}} \widehat{D}(Z_{u,k}) + \tau_u \geq X_u F_{\Lambda_u}^{-1}(1-\epsilon_s), \qquad u \in \mathcal{U}_s,\ s \in \mathcal{S}^{\text{URLLC}},$$
$$\sum_{u\in\mathcal{U}_{s,h}}\sum_{k\in\mathcal{K}} C(Z_{u,k}) + \tau_{s,h} \geq \overline{X}_{s,h} F_{\Lambda_{s,h}}^{-1}(1-\delta_s), \qquad s \in \mathcal{S}^{\text{eMBB}},\ h \in \mathcal{H},$$
$$\sum_{u\in\mathcal{U}_{s,h}}\sum_{k\in\mathcal{K}} C(Z_{u,k}) - \sum_{u\in\mathcal{U}_{s,h}}\sum_{k\in\mathcal{K}} \widehat{D}(Z_{u,k}) + \tau_{s,h} \geq \overline{X}_{s,h} F_{\Lambda_{s,h}}^{-1}(1-\delta_s), \qquad s \in \mathcal{S}^{\text{URLLC}},\ h \in \mathcal{H},$$
$$\sum_{u\in\mathcal{U}_h}\sum_{k\in\mathcal{K}} \widetilde{p}_{u,k} \leq P_h^{\max}, \qquad h \in \mathcal{H},$$
$$\Theta_1(Z_{u,k}, \widetilde{\boldsymbol{p}}) - \widehat{\Theta}_2(Z_{u,k}, \widetilde{\boldsymbol{p}}) + \sigma^2 Z_{u,k} - \tau_{u,k} \leq |g_{u,h,k}|^2\, \widetilde{p}_{u,k}, \qquad u \in \mathcal{U}_h,\ h \in \mathcal{H},\ k \in \mathcal{K},$$
$$Z_{u,k} \geq 0, \qquad u \in \mathcal{U},\ k \in \mathcal{K},$$

where τ, $\tau_{u,k}$, τ_u, and $\tau_{s,h}$, $u \in \mathcal{U}$, $k \in \mathcal{K}$, $s \in \mathcal{S}$, $h \in \mathcal{H}$ are the slack variables for penalizing the objective function. In Line 6, the MNO updates $m_{u,k}^{(i)}$, $\widetilde{p}_{u,k}^{(i)}$, and $Z_{u,k}^{(i)}$ by the optimal solution $m_{u,k}^{*}$, $\widetilde{p}_{u,k}^{*}$, and $Z_{u,k}^{*}$ of problem (21) for user $u \in \mathcal{U}$ and PRB $k \in \mathcal{K}$, respectively. The iteration index is updated in Line 8. The stopping criterion is given in Line 9.

By choosing a small value for $\zeta^{(1)} > 0$ and increasing $\zeta^{(i)}$ in each iteration by a factor η, the optimal solution of problem (21) will converge to a suboptimal solution of the original resource allocation problem (11). In Line 10, the MNO obtains the suboptimal solution $m_{u,k}^{\text{opt}}$, $\widetilde{p}_{u,k}^{\text{opt}}$, and $Z_{u,k}^{\text{opt}}$, $u \in \mathcal{U}$, $k \in \mathcal{K}$ of problem (11). If a feasible solution exists for problem (11), Algorithm 1 finally obtains a solution where $\tau = 0$; $\tau_{u,k} = 0$, $u \in \mathcal{U}$, $k \in \mathcal{K}$; $\tau_u = 0$, $u \in \mathcal{U}$; and $\tau_{s,h} = 0$, $s \in \mathcal{S}$, $h \in \mathcal{H}$ holds for problem (21). Since problem (21) is a convex optimization problem, its optimal solution can be obtained efficiently. Hence, Algorithm 1 converges to a suboptimal solution of problem (11) in polynomial time [16].

IV. Performance Evaluation

In this section, we evaluate the performance of Algorithm 1 via simulations in MATLAB. We consider two network slices, one for eMBB users and another for URLLC users. Unless stated otherwise, the number of users associated with each slice is set to 18. We consider six RRHs placed in a 2×3 grid, where the distance between two adjacent RRHs is set to 50 m. The number of PRBs in the network is set to 25, and the value of B is set to 360 kHz. The wireless channel is modeled by setting the path loss exponent equal to 3.76 and by considering the small-scale Rayleigh fading. The RRH's maximum transmit power $P_h^{\max}$, $h \in \mathcal{H}$ and noise power σ^2 are set to 30 dBm and -114 dBm, respectively. We assume that the average traffic loads of the eMBB and URLLC users are 1 Mbps and 0.1 Mbps, respectively. For slice $s \in \mathcal{S}$ and user $u \in \mathcal{U}$, we set $I_s^{\max} = -74$ dBm, $\alpha_s = 0.5$, and $\beta_u = \frac{1}{18}$. We set $\epsilon_s = 10^{-2}$ and $\delta_s = 10^{-3}$ for eMBB users, and $\epsilon_s = 10^{-3}$ and $\delta_s = 10^{-4}$ for URLLC users. We set $\epsilon_u^{\text{b}} = 10^{-6}$ and $N_u^b = 10$ for the URLLC users. In Algorithm 1, we use hyperparameters $\zeta^{(1)} = 100$, $\zeta^{\max} = 10^5$, and $\eta = 5$. Simulation results are obtained by averaging over 50 simulation runs. For performance comparison, we adopt a baseline scheme with random PRB allocation. Problem (21) is reduced to a transmit power allocation problem that can be solved by the penalized successive convex approximation. Also, when only eMBB users are present in the network, we compare our proposed algorithm with the one proposed in [7].

First, we investigate the impact of the maximum acceptable interference $I_s^{\max}$ on the aggregate throughput of the network. When compared with the baseline scheme, Fig. 2 shows that the aggregate throughput is 30% higher with Algorithm 1 on average. Also, Fig. 2 demonstrates that the aggregate throughput increases with $I_s^{\max}$ and then remains approximately unchanged when $I_s^{\max}$ is larger than -85 dBm. In particular, when $I_s^{\max}$ is small, the MNO has limited freedom to allocate the same PRBs in different RRHs, since the interference constraint (3) can hardly be satisfied. As $I_s^{\max}$ increases,

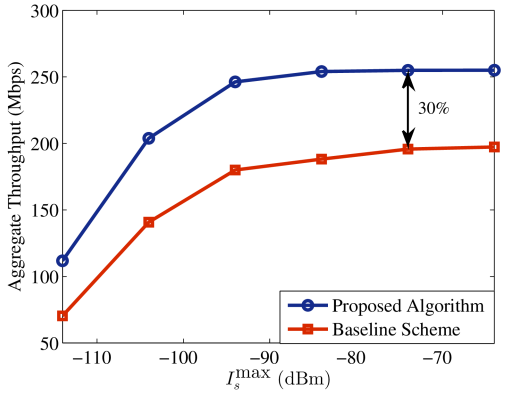

Fig. 2. Aggregate throughput versus the maximum acceptable interference where our proposed algorithm is compared with the baseline scheme.

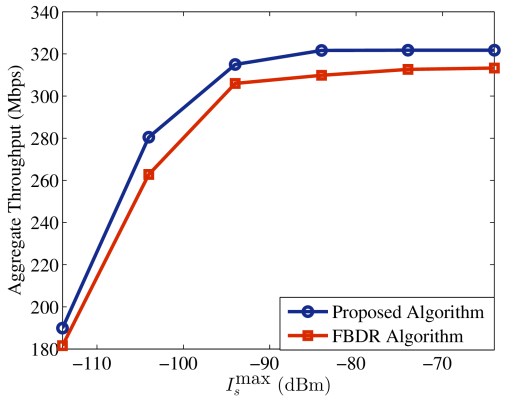

Fig. 3. Aggregate throughput versus the maximum acceptable interference where our proposed algorithm is compared with the FBDR algorithm [7] when only eMBB traffic is considered.

allocating the same PRBs to the users in different RRHs becomes possible, which increases the aggregate throughput of the network. Finally, when $I_s^{\max}$ is greater than -85 dBm, the RRHs cannot impose such interference and the aggregate network throughput remains approximately unchanged due to the fixed allocation of PRBs and transmit power for the users.

In Fig. 3, we compare the performance of Algorithm 1 with the proposed algorithm named fixed BBU capacity and dynamic resource allocation (FBDR) in [7], which applies an iterative subgradient method. The proposed algorithm in [7] assumes that only eMBB users are present in the network. For a fair comparison, in Fig. 3, we only consider the eMBB traffic in Algorithm 1. When compared with the proposed algorithm in [7], Fig. 3 shows that the aggregate throughput is 4% higher with Algorithm 1 on average. Algorithm 1 also converges faster to a suboptimal solution. Since, in this paper, our main focus is on the slicing of eMBB and URLLC services, in the rest of this section, we only compare our proposed algorithm with the baseline scheme.

Next we investigate the aggregate throughput for eMBB and URLLC users when the priority factor α_s for eMBB slice varies from 0.1 to 0.9. Fig. 4 shows that by increasing the priority factor α_s for eMBB slice type, the aggregate throughput for the eMBB users increases, while the aggregate

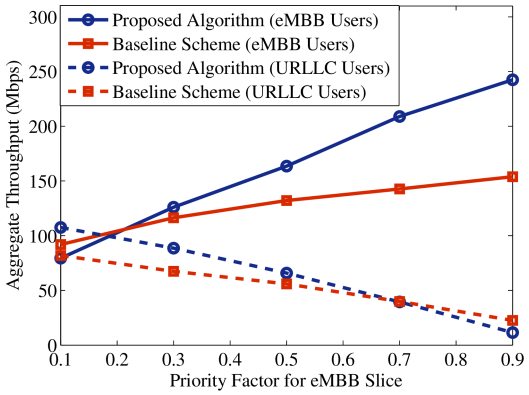

Fig. 4. Aggregate throughput versus priority factor for eMBB slice.

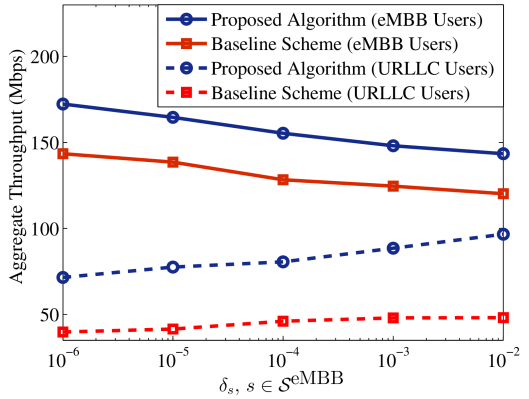

Fig. 5. Aggregate throughput versus the log of the maximum tolerable probability of failure on supporting aggregate traffic load for eMBB users.

throughput for the URLLC users decreases. In particular, assigning a higher priority to the eMBB users causes the MNO to allocate extra PRBs and transmission power to the eMBB users in order to maximize the network aggregate throughput.

Finally, we evaluate the performance of Algorithm 1 for different values of δ_s for the eMBB slice type. We set $\epsilon_s = 0.1$ and $\alpha_s = 0.35$ for the eMBB slice type. We set $\delta_s = 10^{-4}$, $\epsilon_s = 10^{-3}$, and $\alpha_s = 0.65$ for the URLLC slice type. Fig. 5 shows the aggregate throughput for the eMBB and URLLC users, where δ_s for the eMBB slice increases from 10^{-6} to 10^{-2}. When δ_s increases for the eMBB slice type, the aggregate throughput for the eMBB users decreases and the aggregate throughput for the URLLC users increases. The reason is that when δ_s for the eMBB slice type is small, the MNO should reserve more resources for the eMBB users in order to guarantee isolation between slices. Hence, the aggregate throughput for the eMBB users is larger than that of the URLLC users.

V. Conclusion

In this paper, we proposed a resource allocation algorithm for an OFDMA-based C-RAN with multiple RRHs serving users within eMBB and URLLC network slices. By guaranteeing the individual traffic load demand for each user as the QoS constraints and the aggregate traffic load demand as the slice isolation constraints, we formulated the allocation of RRHs' resources (i.e., PRBs, transmit power) as a mixed-integer nonlinear program. The formulated problem took into account the effect of interference between RRHs and short packet transmission for URLLC users. To obtain a suboptimal solution for such an NP-hard optimization problem, we developed an algorithm with polynomial time complexity based on the penalized successive convex approximation. Through simulations, we have shown that compared to a baseline scheme, our proposed algorithm on average achieves 30% higher aggregate network throughput by allocating RRHs' resources among users associated with different RRHs and different network slices. For future work, we plan to consider user admission control in our system model and solve the problem using model-free approaches when the parameters of users' traffic load are unknown.

Acknowledgement

This work was supported by Rogers Communications Canada Inc.

References

[1] K. Katsalis, N. Nikaein, E. Schiller, A. Ksentini, and T. Braun, "Network slices toward 5G communications: Slicing the LTE network," *IEEE Commun. Mag.*, vol. 55, no. 8, pp. 146–154, Aug. 2017.

[2] 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. 55 765–55 779, Sept. 2018.

[3] S. R. Pandey, M. Alsenwi, Y. K. Tun, and C. S. Hong, "A downlink resource scheduling strategy for URLLC traffic," in *Proc. of IEEE Int'l. Conf. on Big Data and Smart Computing*, Kyoto, Japan, Feb. 2019.

[4] M. Alsenwi, N. H. Tran, M. Bennis, A. K. Bairagi, and C. S. Hong, "eMBB-URLLC resource slicing: A risk-sensitive approach," *IEEE Commun. Letters*, vol. 23, no. 4, pp. 740–743, Apr. 2019.

[5] T. Ma, Y. Zhang, F. Wang, D. Wang, and D. Guo, "Slicing resource allocation for eMBB and URLLC in 5G RAN," *Wireless Communications and Mobile Computing*, vol. 2020, Jan. 2020.

[6] Y. Hua, R. Li, Z. Zhao, X. Chen, and H. Zhang, "GAN-powered deep distributional reinforcement learning for resource management in network slicing," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 2, pp. 334–349, Feb. 2020.

[7] Y. L. Lee, J. Loo, T. C. Chuah, and L.-C. Wang, "Dynamic network slicing for multitenant heterogeneous cloud radio access networks," *IEEE Trans. on Wireless Commun.*, vol. 17, no. 4, pp. 2146–2161, Apr. 2018.

[8] C. Marquez, M. Gramaglia, M. Fiore, A. Banchs, and X. Costa-Perez, "How should I slice my network? A multi-service empirical evaluation of resource sharing efficiency," in *Proc. of ACM Int'l. Conf. on Mobile Computing and Networking (MobiCom)*, New Delhi, India, Oct. 2018.

[9] J. Scarlett, V. Y. Tan, and G. Durisi, "The dispersion of nearest-neighbor decoding for additive non-Gaussian channels," *IEEE Trans. on Information Theory*, vol. 63, no. 1, pp. 81–92, Jan. 2017.

[10] 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 Commun. Surveys & Tutorials*, vol. 22, no. 2, pp. 905–929, 2nd quarter 2020.

[11] 3GPP TR 36.889 V13.0.0, "Study on Licensed-Assisted Access to Unlicensed Spectrum; (Release 13)," Jun. 2015.

[12] F. A. Haight, *Handbook of the Poisson Distribution*. Wiley, 1967.

[13] D. Bega, M. Gramaglia, A. Banchs, V. Sciancalepore, K. Samdanis, and X. Costa-Perez, "Optimising 5G infrastructure markets: The business of network slicing," in *Proc. of IEEE INFOCOM*, Atlanta, GA, May 2017.

[14] J. Lee and S. Leyffer, *Mixed Integer Nonlinear Programming*. Springer, 2011.

[15] H. Tuy, *Convex Analysis and Global Optimization*. Springer, 2016.

[16] W. R. Ghanem, V. Jamali, Y. Sun, and R. Schober, "Resource allocation for multi-user downlink MISO OFDMA-URLLC systems," *arXiv preprint arXiv:1910.06127*, Oct. 2019.