Incorporating URLLC and Multicast eMBB in Sliced Cloud Radio Access Network

Jianhua Tang[†], Byonghyo Shim[†], Tsung-Hui Chang[*] and Tony Q. S. Quek[‡]
[†] INMC, Department of Electrical and Computer Engineering, Seoul National University, Korea
[*] School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, China
[‡] Singapore University of Technology and Design, Singapore
jianhua_tang@islab.snu.ac.kr, bshim@snu.ac.kr, tsunghui.chang@ieee.org, tonyquek@sutd.edu.sg

***Abstract*—The fifth generation (5G) wireless systems aims to differentiate its services based on different application scenarios. Instead of constructing different physical networks to support each application, radio access network (RAN) slicing is deemed as a prospective solution to help operate multiple logical separated wireless networks in a single physical network. In this paper, we incorporate two typical 5G services, i.e., enhanced Mobile BroadBand (eMBB) and Ultra-Reliable Low-Latency Communications (URLLC), in a cloud RAN (C-RAN), which is suitable for RAN slicing due to its high flexibility. In particular, for eMBB, we make use of multicasting to improve the throughput, and for URLLC, we leverage finite blocklength capacity to capture the delay accurately. Our objective is to minimize the total power consumption, subject to the limited physical resource constraints. We formulate the problem as a nonconvex optimization problem and exploit efficient approaches to solve it, such as successive convex approximation and semidefinite relaxation. Simulation results show that our proposed algorithm saves system power consumption significantly.**

***Index Terms*—URLLC, eMBB, multicast, C-RAN, network slicing**

I. Introduction

The services catered by the incoming fifth generation (5G) wireless systems is expected to fall into three categories [2], [3], i.e., enhanced Mobile BroadBand (eMBB), Ultra-Reliable Low-Latency Communications (URLLC), and massive Machine-Type Communications (mMTC). Specifically, eMBB requires high data rate and reliable broadband access over large areas, URLLC supports ultra-low latency transmission for small payload with a high level of reliability, and mMTC is targeted for the wireless connectivity for massive number of sporadically active Internet of Things (IoT) devices. Majority of works to date are individual for these three services [4]–[6]. Nevertheless, how to efficiently and simultaneously support them in a shared physical system is still an unaddressed problem.

Recently, network slicing [7], [8] has received much attention as a promising technique to provide flexibility and scalability for a variety of 5G services. The main feature of network slicing is to run multiple logical separated networks as independent business operations on top of a common shared physical infrastructure [9]. With network slicing, network resources can be elastically and dynamically allocated to the logical network slices according to on-demand tailored service requirements. Thus, it offers the hope to resolve the aforementioned unaddressed problem. Whereas, to facilitate network slicing, an agile and programmable physical network architecture is a requisite.

Cloud radio access network (C-RAN) has emerged as a prospective architecture for 5G. A typical structure of C-RAN includes three main components: remote radio heads (RRHs), fronthaul links, and baseband unit (BBU) pool. The most significant innovation part of C-RAN is that it decouples baseband processing functionalities from the RRHs and migrates these functionalities to the centralized cloud BBU pool, which consists of many general-purpose servers. With the centralized cloud BBU pool, an agile and programmable software-defined environment in the RAN side is now achieved [10].

In this work, with the merits from network slicing and C-RAN, we attempt to incorporate both eMBB and URLLC services in C-RAN. That is, two different types of network slices are tailored in C-RAN to support these two different services. Particularly, we consider multicast transmission for the eMBB slice, since multicast transmission is envisioned to be a popular transmission scheme in many 5G scenarios [11]. For example, in C-RAN with Caching as a Service [12], popular contents are stored in the centralized BBU pool. This centralized content caching structure is easy to facilitate multicast transmission among RRHs to deliver data to a user group with the same interest. Another application example is live video streaming for some hot events (e.g. FIFA World Cup) where the video stream goes from stream server to UEs via the centralized BBU pool.

There have been many previous efforts on network slicing. Most of them focus on the problems in the upper layers of network, such as resource orchestration [13] and service chaining [14]. Compared to the upper layers network slicing, there is one more difficulty in RAN slicing, i.e., *inter-slice*

The full version of this paper can be found in [1]. This work was supported in part by the Korea Research Fellowship Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (Grant No. 2016H1D3A1938245), the NRF grant funded by the Korean government (MSIP) (2014R1A5A1011478), the framework of international cooperation program managed by NRF (2016K1A3A1A20006019), the NSFC, China, under Grant 61571385 and 61731018, and by the Shenzhen Fundamental Research Fund under Grant ZDSYS201707251409055 and KQTD2015033114415 45, and the SUTD-ZJU Research Collaboration under Grant SUTD-ZJU/RES/01/2016 and SUTD-ZJU/RES/05/2016.

978-1-5386-8088-9/19/$31.00 ©2019 IEEE

interference isolation. Unlike the upper layers network slicing, whose resources in different slices can be isolated by virtualization, different slices share the same physical channel in RAN slicing. Real challenge in this case is how to isolate the interference between different slices is a challenge.

An aim of this paper is to propose a novel scheme to incorporate eMBB and URLLC slices in C-RAN to minimize total power consumption under both inter-slice (e.g, total bandwidth of the system) and intra-slice constraints (e.g, quality-of-service (QoS) to each user), while guarantee the inter-slice interference isolation as well.

A. Our contributions

Our main contributions are as follows:

1) We tame the following interesting tussles in our system model:
   - **Multicast vs. unicast:** The eMBB slices use multicast transmission while the URLLC slices still rely on unicast transmission.
   - **High throughput vs. low delay:** The eMBB slice aims to have a high throughput, while the URLLC slice needs a low latency per packet.
   - **Shannon's capacity vs. finite blocklength capacity:** The achievable rate of the eMBB slice can be captured by Shannon's capacity, while the counterpart of URLLC slice depends on finite blocklength capacity [15], [16] due to the small payload.
2) We leverage efficient methods to solve the intractable power minimization problem, such as semidefinite relaxation (SDR) and successive convex approximation (SCA). We also prove the tightness of SDR under certain cases. Our simulation results verify the efficiency of our proposed approach.

Notations

We use calligraphy letters to represent the sets, boldface lower case letters to denote the vectors, and boldface upper case letters to denote the matrices. $\|\mathbf{x}\|_2$ and $\|\mathbf{X}\|_F$ stand for the Euclidean norm and Frobenius norm respectively. $(\cdot)^H$ represents the conjugate transpose. $\mathbf{X} \succeq 0$ means that matrix $\mathbf{X}$ is Hermitian positive semidefinite. $\mathbb{C}$ and $\mathbb{R}$ stand for complex numbers and real numbers respectively. The $\log(\cdot)$ function is the logarithm function with base 2.

II. SYSTEM MODEL AND PROBLEM FORMULATION

In this work, we consider two different network slices, i.e., one multicast eMBB slice and one URLLC slice. We use the tuple $\{I^c,\ R\}$ to represent the multicast eMBB slice, in which I^c denotes the number of users in the multicast eMBB slice and R indicates the minimum throughput requirement. Also, we use $\{I^u,\ D\}$ to represent the URLLC slice, where I^u denotes the number of users in the URLLC slice and R indicates the maximum packet delay threshold.

To achieve the inter-slice interference isolation and also adaptively guarantee the quality-of-service (QoS) for each slice, we leverage the flexible frequency division duplex (FDD) technique. In this scheme, the frequency resource size assigned to each user can be tailored. This concept is similar to the flexible radio framing [17], [18], which is the evolution of the celebrated orthogonal frequency-division multiple access (OFDMA) framing. In flexible radio framing, the transmission time interval (TTI) size can be dynamically adjusted. This provides the flexibility to achieve low latency communication. In this paper, we assign b^c bandwidth to the multicast eMBB slice and b^u bandwidth to the URLLC slice. We denote the total bandwidth of the system as B.

We consider the data sharing transmission in the downlink C-RAN, where each user equipments (UE's) desired data can be shared among all the coordinated RRHs. Suppose there are J coordinated RRHs, each with K antennas. We denote the set of all coordinated RRHs as $\mathcal{J} = \{1, \cdots, J\}$, UEs under eMBB slice as $\mathcal{I}^c = \{1, \cdots, I^c\}$, and UEs under URLLC slice as $\mathcal{I}^u = \{1, \cdots, I^u\}$. We assume that each UE is equipped with single receive antenna. The channel from RRH j to UE i is denoted as $\mathbf{h}_{ij}$, where $\mathbf{h}_{ij} \in \mathbb{C}^K$, for $i \in \mathcal{I}^c \bigcup \mathcal{I}^u$ and $j \in \mathcal{J}$. Suppose that $\mathbf{h}_{ij}$ is drawn from a certain random distribution, and this distribution is known in advance by the C-RAN operator. The random variables $\mathbf{h}_{ij}$, for any $i \in \mathcal{I}^c \bigcup \mathcal{I}^u$ and $j \in \mathcal{J}$, are independent and identically distributed.

A. Multicast eMBB slice

In this work, we only consider the single-group multicasting [19]. In this scheme, each eMBB slice serves a group of users having the same interest. For the multicast group, let u^c be the data symbol to all UEs in this group with $\mathbb{E}[|u^c|^2] = 1$, and $\mathbf{v}_j \in \mathbb{C}^K$ be the transmit beamformer to all UEs in this group from RRH j.

Then the received signal at UE i in multicast eMBB slice is,

$$\hat{u}_i = \sum_{j \in \mathcal{J}} \mathbf{h}_{ij}^H \mathbf{v}_j u^c + \delta_i,\ \forall i \in \mathcal{I}^c,$$

where the first term is the desired signal for UE i and $\delta_i \sim \mathcal{CN}(0, \sigma_i^2)$ is the additive white Gaussian noise (AWGN) at UE i. The corresponding signal-to-noise ratio (SNR) at UE i is

$$\mathrm{SNR}_i^c = \frac{|\sum_{j \in \mathcal{J}} \mathbf{h}_{ij}^H \mathbf{v}_j|^2}{\sigma_i^2}. \tag{1}$$

Then, the achievable rate of this multicast group is

$$r^c \leq \min_{i \in \mathcal{I}^c} \{\log(1 + \mathrm{SNR}_i^c)\}. \tag{2}$$

The corresponding throughput requirement is

$$r^c b^c \geq R. \tag{3}$$

B. URLLC slice

For the URLLC slice, let u_i^u be the data symbol for the i-th UE with $\mathbb{E}[|u_i^u|^2] = 1$, and $\mathbf{w}_{ij} \in \mathbb{C}^K$ be the transmit beamformer to UE i from RRH j . Suppose that the flexible FDD is applied inside the URLLC slice. Hence, at UE i, there is no interference from other UEs.

On this basis, the received signal at UE $i \in \mathcal{I}^u$ is

$$\bar{u}_i = \sum_{j \in \mathcal{J}} \mathbf{h}_{ij}^H \mathbf{w}_{ij} u_i^u + \delta_i.$$

The corresponding SNR at UE $i \in \mathcal{I}^u$ is

$$\text{SNR}_i^u = \frac{|\sum_{j \in \mathcal{J}} \mathbf{h}_{ij}^H \mathbf{w}_{ij}|^2}{\sigma_i^2}.$$

In URLLC, packets are typically very short, so that the achievable rate and the transmission error probability cannot be accurately captured by Shannon's capacity. Instead, the achievable rate in URLLC falls in the finite blocklength channel coding regime (see [15] for details). Let r_i^u be the achievable rate of UE i in the URLLC slice and n_i be the length of codeword block (in symbols). Then we have [15]

$$r_i^u \leq \log(1 + \text{SNR}_i^u) - \sqrt{\frac{C_i}{n_i}} Q^{-1}(\epsilon) \log e, \ \forall i \in \mathcal{I}^u, \quad (4)$$

where $Q^{-1}(\cdot)$ is the inverse of the Gaussian Q-function, $\epsilon > 0$ is the transmission error probability, and C_i is the *channel dispersion*[1] of UE i, given by

$$C_i = 1 - \frac{1}{(1 + \text{SNR}_i^u)^2}. \quad (5)$$

Let F_i be the packet size to UE i. Then, the delay constraint for UEs in the URLLC slice is

$$\max_{i \in \mathcal{I}^u} \frac{F_i}{r_i^u b_i^u} \leq D, \quad (6)$$

where b_i^u is the bandwidth assigned to UE i, such that $\sum_{i \in \mathcal{I}^u} b_i^u \leq b^u$.

C. Inter-slice constraints

Since each RRH has its maximum transmitting power E_j constraint, we have

$$\mathbf{v}_j^H \mathbf{v}_j + \sum_{i \in \mathcal{I}^u} \mathbf{w}_{ij}^H \mathbf{w}_{ij} \leq E_j, \ \forall j \in \mathcal{J}. \quad (7)$$

In addition, the bandwidth resources allocated to these two slices have to satisfy

$$b^c + \sum_{i \in \mathcal{I}^u} b_i^u \leq B. \quad (8)$$

D. Problem formulation

Let $\mathbf{v} = [\mathbf{v}_1 \ \mathbf{v}_2 \cdots \mathbf{v}_j]^T \in \mathbb{C}^{JK \times 1}$ and $\mathbf{w}_i = [\mathbf{w}_{i1} \ \mathbf{w}_{i2} \cdots \mathbf{w}_{ij}]^T \in \mathbb{C}^{JK \times 1}$. The objective of this work is to minimize the total power consumption. And thus, the problem can be formulated as

$$\begin{aligned}
\text{(P0)} \quad \min_{\substack{b^c,\ b_i^u,\\ \mathbf{v},\ \mathbf{w}_i}} \quad & \sum_{j \in \mathcal{J}} \mathbf{v}_j^H \mathbf{v}_j + \sum_{j \in \mathcal{J}} \sum_{i \in \mathcal{I}^u} \mathbf{w}_{ij}^H \mathbf{w}_{ij} \\
\text{s.t.} \quad & b^c + \sum_{i \in \mathcal{I}^u} b_i^u \leq B, \\
& r^c \leq \min_{i \in \mathcal{I}^c} \{\log(1 + \text{SNR}_i^c)\}, & (9) \\
& r^c b^c \geq R, & (10) \\
& r_i^u \leq \log(1 + \text{SNR}_i^u) \\
& \quad - \sqrt{\frac{C_i}{n_i}} Q^{-1}(\epsilon) \log e, \ \forall i \in \mathcal{I}^u, & (11) \\
& \frac{F_i}{r_i^u b_i^u} \leq D, \ \forall i \in \mathcal{I}^u, & (12) \\
& \mathbf{v}_j^H \mathbf{v}_j + \sum_{i \in \mathcal{I}^u} \mathbf{w}_{ij}^H \mathbf{w}_{ij} \leq E_j, \ \forall j \in \mathcal{J}. & (13)
\end{aligned}$$

In problem (P0), constraints (2), (3), (4), and (6) are noncovex, which further complicate problem (P0). In the following sections, we propose an approach to obtain an approximate solution for problem (P0).

III. Solution for Two Categories

In this section, we first solve the nonconvex problem (P0) under the high SNR regime and then propose the solution approach under general SNR case.

Lemma 1: In problem (P0), constraints (10) and (12) are active inequality constraints. That is, an optimal solution $\{b^c, \ b_i^u, \ \mathbf{v}, \ \mathbf{w}_i\}$ to problem (P0) satisfies

$$\begin{cases} r^c b^c = R, \\ r_i^u b_i^u = \dfrac{F_i}{D}. \end{cases} \quad (14)$$

Proof: Suppose that there is an optimal solution $\{b^c, \ b_i^u, \ \mathbf{v}, \ \mathbf{w}_i\}$ satisfying $r^c b^c > R$ or $r_i^u b_i^u > F_i/D$. This implies that we can still scale down $\mathbf{v}$ or $\mathbf{w}_i$ in the feasible region to increase the objective, which contradicts the optimality and establish the lemma. □

Based on Lemma 1, we use $\frac{R}{b^c}$ and $\frac{F_i}{D b_i^u}$ to replace r^c and r_i^u in (10) and (12) respectively. Further, let $\mathbf{V} = \mathbf{v}\mathbf{v}^H \in \mathbb{R}^{JK \times JK}$, $\mathbf{W}_i = \mathbf{w}_i \mathbf{w}_i^H \in \mathbb{R}^{JK \times JK}$ and $\mathbf{H}_i = \mathbf{h}_i \mathbf{h}_i^H \in \mathbb{R}^{JK \times JK}$, where $\mathbf{h}_i = [\mathbf{h}_{i1} \ \mathbf{h}_{i2} \cdots \mathbf{h}_{ij}]^T \in \mathbb{C}^{JK \times 1}$. We can reformulate constraints (10), (12) and (13) as

$$\begin{aligned}
& \frac{R}{b^c} \leq \log(1 + \text{tr}\,(\mathbf{H}_i \mathbf{V})/\sigma_i^2), \ \forall i \in \mathcal{I}^c, & (15) \\
& \frac{F_i}{b_i^u D} \leq \log(1 + \text{tr}\,(\mathbf{H}_i \mathbf{W}_i)/\sigma_i^2) \\
& \qquad - \sqrt{1 - \frac{1}{(1 + \text{tr}\,(\mathbf{H}_i \mathbf{W}_i)/\sigma_i^2)^2}} \frac{Q^{-1}(\epsilon) \log e}{\sqrt{n_i}}, \\
& \qquad \forall i \in \mathcal{I}^u, & (16) \\
& E_j \geq \text{tr}\,(\mathbf{G}_j \mathbf{V}) + \sum_{i \in \mathcal{I}^u} \text{tr}\,(\mathbf{G}_j \mathbf{W}_i), \forall j \in \mathcal{J}, & (17)
\end{aligned}$$

where $\mathbf{G}_j$ is a square matrix with $J \times J$ blocks, and each block in $\mathbf{G}_j$ is a $K \times K$ matrix. In $\mathbf{G}_j$, the block in the j-th row and j-th column is a $K \times K$ identity matrix, and all other blocks are zero matrices. Then, by applying the following property

$$\begin{cases} \mathbf{V} = \mathbf{v}\mathbf{v}^H \Leftrightarrow \mathbf{V} \succeq 0, \ \text{rank}(\mathbf{V}) \leq 1, \\ \mathbf{W}_i = \mathbf{w}_i \mathbf{w}_i^H \Leftrightarrow \mathbf{W}_i \succeq 0, \ \text{rank}(\mathbf{W}_i) \leq 1, \end{cases}$$

[1] Other than in [15], here we put $\log e$ in (4) for simplicity.

we can obtain an equivalent formulation of problem (P0) as

$$\begin{aligned} \text{(P-S1)} \quad \min_{b^c,\ b_i^u,\ \mathbf{V},\ \mathbf{W}_i} \quad & \operatorname{tr}(\mathbf{V}) + \sum_{i\in\mathcal{I}^u} \operatorname{tr}(\mathbf{W}_i) \\ \text{s.t.} \quad & (8),\ (15),\ (16),\ (17), \\ & \mathbf{V} \succeq 0, && (18) \\ & \mathbf{W}_i \succeq 0,\ \forall i \in \mathcal{I}^u, && (19) \\ & \operatorname{rank}(\mathbf{V}) \le 1, && (20) \\ & \operatorname{rank}(\mathbf{W}_i) \le 1,\ \forall i \in \mathcal{I}^u. && (21) \end{aligned}$$

Using Lemma 1 and also after the change of variables, problem (P-S1) is simplified from problem (P0). However, problem (P-S1) is still very difficult to handle due to the nonconvexity of constraints (16), (20) and (21). In the following subsections, we leverage the following approaches to resolve the difficulties.

- For the rank constraints, we resort to the semidefinite relaxation (SDR) method. Specifically, we drop the rank constraints first, and then solve the problem without rank constraints. If the resulting $\mathbf{V}$ and $\mathbf{W}_i$ are of rank one or zero, we conclude that the SDR is tight and no more manipulation is needed [20]. On the other hand, if the rank of resulting $\mathbf{V}$ or $\mathbf{W}_i$ is larger than one, we must use an approach to extract the approximate solution from it, e.g., the randomization method [21].
- For constraint (16), we start from the high SNR regime first. It is verified in [22] that, under high SNR regime, the channel dispersion in (4) approaches to 1. Hence, constraint (16) can be greatly simplified under high SNR regime. After the discussion of high SNR regime, we then propose an approximation approach for the general SNR case.

A. Solution for URLLC slice under high SNR regime

Let τ be the high SNR threshold. Based on the results in [22], $C_i \approx 1$, for $\mathrm{SNR}_i^u \ge \tau$. Therefore, constraint (16) is simplified as

$$\frac{F_i}{b_i^u D} \le \log(1 + \operatorname{tr}(\mathbf{H}_i\mathbf{W}_i)/\sigma_i^2) - \frac{Q^{-1}(\epsilon)\log e}{\sqrt{n_i}},\ \forall i \in \mathcal{I}^u. \tag{22}$$

Applying SDR to problem (P-S1), we get the following problem

$$\begin{aligned} \text{(P-S2)} \quad \min_{b^c,\ b_i^u,\ \mathbf{V},\ \mathbf{W}_i} \quad & \operatorname{tr}(\mathbf{V}) + \sum_{i\in\mathcal{I}^u} \operatorname{tr}(\mathbf{W}_i) \\ \text{s.t.} \quad & (8),\ (15),\ (17),\ (18),\ (19),\ (22), \\ & \operatorname{tr}(\mathbf{H}_i\mathbf{W}_i)/\sigma_i^2 \ge \tau,\ \forall i \in \mathcal{I}^u. && (23) \end{aligned}$$

Problem (P-S2) is a convex optimization problem and can be easily solved by the interior point method, which has been always implemented in standard optimization toolboxes, e.g. CVX [23]. If we denote $\mathbf{V}^*$ and $\mathbf{W}_i^*$ as the optimal solution of $\mathbf{V}$ and $\mathbf{W}_i$ in problem (P-S2) respectively, then the following theorem shows the effectiveness of utilizing SDR in problem (P-S2).

Theorem 1: In problem (P-S2), the SDR for $\mathbf{W}_i$ is tight. That is,

$$\operatorname{rank}(\mathbf{W}_i^*) \le 1,\ \forall i \in \mathcal{I}^u.$$

However, the SDR for $\mathbf{V}$ may not be tight.

Proof: See Appendix A. □

B. Solution for URLLC slice under general SNR

Recall the channel dispersion in (16),

$$C_i = 1 - \frac{1}{\alpha_i^2},\ \forall i \in \mathcal{I}^u, \tag{24}$$

where $\alpha_i = 1 + \operatorname{tr}(\mathbf{H}_i\mathbf{W}_i)/\sigma_i^2$. It can be verified that $\sqrt{C_i}$ is concave with respect to (w.r.t.) $\alpha_i > 1$. On this basis, we can employ successive convex approximation (SCA) to tackle the nonconvex constraint (16). More details are given below.

SCA is an efficient way to solve various types of nonconvex optimization problems [24], [25]. The main idea of SCA is that, a locally tight approximation of the original problem is performed at each iteration to produce a tight convex objective function and constraint sets. In other words, instead of solving a nonconvex optimization problem directly, a series of convex optimization problem is solved iteratively to obtain an approximate solution. In this paper, we utilize the approximation functions to locally approximate the nonconvex functions based on the following assumption [25], [26].

Assumption 1: A function $\tilde{h}(x,y)$ is called as the *approximation function* for the nonconvex function $h(x)$, when the following conditions hold:

- $\tilde{h}(x,y)$ is continuous in (x,y).
- $\tilde{h}(x,y)$ is convex in x.
- The function value of $\tilde{h}(x,x)$ and $h(x)$ are consistent, i.e, $\tilde{h}(x,x) = h(x),\ \forall x$.
- The gradient $\frac{\partial \tilde{h}(x,y)}{\partial x}|_{x=y}$ and $\nabla h(x)|_{x=y}$ are consistent, i.e., $\frac{\partial \tilde{h}(x,y)}{\partial x}|_{x=y} = \nabla h(x)|_{x=y},\ \forall x$.
- $\tilde{h}(x,y)$ is an upper-bound of $h(x)$, i.e., $\tilde{h}(x,y) \ge h(x),\ \forall x,\ y$.

Applying SCA to $\sqrt{C_i}$, at iteration p, we have

$$\begin{aligned} \sqrt{C_i} \le & \sqrt{1 - \frac{1}{(\alpha_i^{(p-1)})^2}} \\ & + \beta_i^{(p-1)}\left(1 + \operatorname{tr}(\mathbf{H}_i\mathbf{W}_i)/\sigma_i^2 - \alpha_i^{(p-1)}\right), \end{aligned} \tag{25}$$

where $\alpha_i^{(p-1)} = 1 + \operatorname{tr}\left(\mathbf{H}_i\mathbf{W}_i^{(p-1)}\right)/\sigma_i^2$ and $\beta_i^{(p-1)} = (\alpha_i^{(p-1)})^{-2}((\alpha_i^{(p-1)})^2 - 1)^{-0.5}$ are constants obtained from the $(p-1)$-th iteration. It can be verified that the approximation in (25) satisfies Assumption 1. Thus, at iteration p, constraint (16) can be approximated as

$$\begin{aligned} \frac{F_i}{b_i^u D} \le & \log\left(1 + \frac{\operatorname{tr}(\mathbf{H}_i\mathbf{W}_i)}{\sigma_i^2}\right) - \left(\sqrt{1 - \frac{1}{(\alpha_i^{(p-1)})^2}}\right. \\ & \left. + \beta_i^{(p-1)}\left(1 + \frac{\operatorname{tr}(\mathbf{H}_i\mathbf{W}_i)}{\sigma_i^2} - \alpha_i^{(p-1)}\right)\right) \frac{Q^{-1}(\epsilon)\log e}{\sqrt{n_i}}, \\ & \forall i \in \mathcal{I}^u,\ \forall s \in \mathcal{S}^{u+}, \end{aligned} \tag{26}$$

which is a convex constraint.

Hence, for the general SNR scenario, if we employ SDR on problem (P-S1), the following problem is required to be solved at iteration p:

$$\text{(P-S3)} \quad \min_{b^c,\ b_i^u,\ \mathbf{V},\ \mathbf{W}_i} \quad \operatorname{tr}(\mathbf{V}) + \sum_{i \in \mathcal{I}^u} \operatorname{tr}(\mathbf{W}_i)$$
$$\text{s.t.} \quad (8),\ (15),\ (17),\ (18),\ (19),\ \text{and } (26).$$

Problem (P-S3) is a convex optimization problem and can be resolved by standard tools as well.

Let $\left\{b^{c(p)},\ b_i^{u(p)},\ \mathbf{V}^{(p)},\ \mathbf{W}_i^{(p)}\right\}$ be the optimal solution for problem (P-S3) at the p-th iteration. We elaborate the SCA+SDR algorithm for problem (P-S1) under the general SNR in Algorithm 1, in which $O^{(p)}$ is the optimal objective function value of problem (P-S1) at iteration p and $\varrho > 0$ is a small constant.

Algorithm 1 SCA+SDR algorithm to solve problem (P-S1)

```
1: Initialization: $\left\{b^{c(0)},\ b_i^{u(0)},\ \mathbf{V}^{(0)},\ \mathbf{W}_i^{(0)}\right\}$.
2: Iteration $p \geq 1$: Solving problem (P-S3) with given $\left\{b^{c(p-1)},\ b_i^{u(p-1)},\ \mathbf{V}^{(p-1)},\ \mathbf{W}_i^{(p-1)}\right\}$, and obtain $\left\{b^{c(p)},\ b_i^{u(p)},\ \mathbf{V}^{(p)},\ \mathbf{W}_i^{(p)}\right\}$.
3: if $|O^{(p)} - O^{(p-1)}| < \varrho$ then
4:     Problem (P-S1) achieves the approximated solution, stop iteration;
5: else
6:     Let $p = p + 1$, go to step 2.
7: end if
8: Output: $\left\{b^{c(p)},\ b_i^{u(p)},\ \mathbf{V}^{(p)},\ \mathbf{W}_i^{(p)}\right\}$.
```

The following theorem unravels the convergence of Algorithm 1 and the tightness of SDR for $\mathbf{W}_i^{(p)}$ under general SNR (for the URLLC slice).

Proposition 1: Every limit point $\mathbf{W}_i^{(\infty)}$ generated by Algorithm 1 is a stationary point of problem (P-S1). That is,

$$\lim_{p\to\infty} \left\| \mathbf{W}_i^{(p)} - \mathbf{W}_i^{(p-1)} \right\|_F = 0,\ \forall i \in \mathcal{I}^u.$$

Furthermore, if the Slater condition holds at the limit point $\mathbf{W}_i^{(\infty)}$, then

1) $\mathbf{W}_i^{(\infty)}$ is a KKT point;
2) The SDR for $\mathbf{W}_i^{(p)}$ is asymptotically tight. That is,

$$\lim_{p\to\infty} \operatorname{rank}\left(\mathbf{W}_i^{(p)}\right) \leq 1,\ \forall i \in \mathcal{I}^u.$$

Proof: A similar proof can be found in Appendix A of [26], we omit it for brevity. □

IV. Simulation results

We consider a C-RAN with 3 RRHs, which are located on a circle with radius 0.5 km. The distances between each two RRHs are equal. UEs from different slices are randomly, uniformly and independently distributed within this disk. The received power at a UE located d km away from a RRH is given by p (dB) $= 128.1 + 37.6 \log_{10} d$. The transmit antenna gain at each RRH is 5 dB. The lognormal shadowing parameter is set to 10 dB. In our simulations, we consider homogeneous RRHs with $E_j = E, \forall j$, and homogeneous UEs with $\sigma_i^2 = \sigma^2$, $F_i = F$, and $n_i = n$, $\forall i$. Our default simulation parameters [27] are summarized in Table I.

TABLE I
Simulation parameters

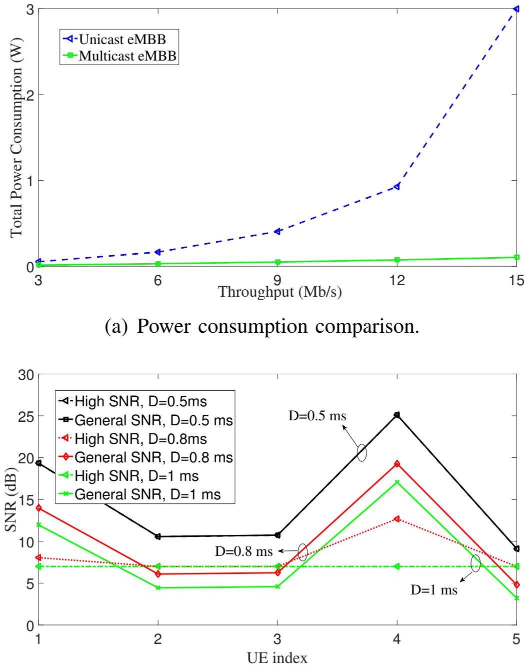
Parameter	Value	Parameter	Value
J	3	K	2
E	1 W	F	500 bytes
σ^2	-83.98 dBm/Hz	B	10 MHz
n	168	τ	7 dB

A. Single slice results

To comprehensively understand the effectiveness of our proposed model and algorithm, we first show the performance result when the system only supports one slice, i.e., one eMBB slice or one URLLC slice respectively.

Firstly, we assume that the system only supports one eMBB slice. We are interested in examining the system power consumption when this eMBB slice uses difference schemes, i.e., unicast or multicast respectively. To avoid interference between different UEs in unicast eMBB, we also apply FDD (as we did for URLLC slice). In Fig. 1(a), we show the system power consumption under different throughput requirements R (there are 4 UEs in this slice). From Fig. 1(a), we can observe that the system power consumption of unicast eMBB is much higher than multicast eMBB. When the throughput requirement increases, system power consumption of unicast eMBB goes up almost exponentially, while that for the multicast eMBB grows up moderately. That is because, under multicast scheme, every UE makes full use of the whole bandwidth B. In contrast, under unicast scheme, every UE just partially use the bandwidth, then the RRH side has to spend more transmit power on each UE to satisfy the throughput requirement.

Secondly, we assume that the system only supports one URLLC slice. We are interested in investigating the system power consumption and SNR values when apply different solution approaches, i.e., high SNR regime solution (in Section III-A) or general SNR regime solution (in Section III-B) respectively. We show the result under three different delay requirements in Fig. 1(b), i.e., $D = 0.5$ ms, 0.8 ms and 1 ms respectively, under the high SNR threshold $\tau = 7$ dB. We observe from Fig. 1(b) that, for the general SNR solution approach, the optimal SNR value for each UE can be either larger or smaller than τ when delay requirements are not very stringent, i.e., $D = 0.8$ ms and 1 ms. However, when the delay requirement becomes stringent, i.e., $D = 0.5$ ms, every UE's SNR should be larger than τ, and the curves of high SNR and general SNR overlap. In addition, we also show the system power consumption from two different solution approaches in Table II. It can be concluded that, compared to the high SNR solution approach, the general SNR solution approach saves



(a) Power consumption comparison.

(b) SNR values for UEs in URLLC slice.

Fig. 1. Single slice results.

TABLE II
POWER CONSUMPTION FOR URLLC SLICE

	High SNR	General SNR
$D = 0.5$ ms	0.85 W	0.85 W
$D = 0.8$ ms	0.38 W	0.30 W
$D = 1$ ms	0.38 W	0.20 W

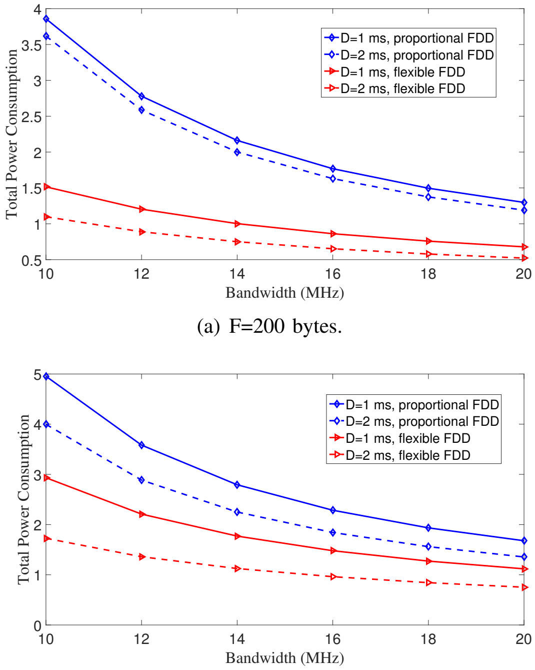
much power consumption when delay requirements are not very stringent.

B. Two slices results

In this subsection, we incorporate two slices in the system, i.e., one multicast eMBB slice with 4 UEs and one URLLC slice with 3 UEs. All results in this subsections are generated based on the general SNR solution (for URLLC slice).

In this work, we use flexible FDD to isolate the interference between two slices, and the bandwidth allocated for each slice is calculated by solving problem (P0). In fact, a simple bandwidth allocation scheme can also isolate the inter-slice interference, i.e., *proportional FDD*. In proportional FDD, we segment the total bandwidth into $I^u + 1$ equivalent blocks. Every UE in the URLLC slice occupies one individual block, and all UEs in the multicast eMBB slice share the one remaining block.

We compare the flexible FDD and the proportional FDD in Fig. 2 w.r.t. the total power consumptions. We can learn from Fig. 2 that our proposed flexible FDD scheme saves power consumption significantly. And power consumption increases with larger packet size and more stringent delay requirement.



(a) F=200 bytes.

(b) F=500 bytes.

Fig. 2. Two slices results.

V. CONCLUSION

In this paper, we incorporated both URLLC and multicast eMBB slices in C-RAN. We used flexible FDD to isolate the inter-slice interference. We applied efficient approaches, such as SCA and SDR, to solve the power consumption minimization problem. The performance advances of our proposed algorithm w.r.t. power consumption were examined by comprehensive simulations.

As a future work, we will consider non-orthogonal slicing (which means two different slices may interfere each other), to further improve the resource utilization in our problem.

APPENDIX A
PROOF OF THEOREM 1

The Lagrangian for problem (P-S2) is (we only include the terms that relevant to this proof),

$$\begin{aligned}\mathscr{L} =& \text{tr}\left(\mathbf{V}\right) + \sum_{i\in\mathcal{I}^u} \text{tr}\left(\mathbf{W}_i\right) - \sum_{i\in\mathcal{I}^c} \mu_i \left(\log(1+\text{tr}\left(\mathbf{H}_i\mathbf{V}\right)/\sigma_i^2)\right) \\ &+ \sum_{j\in\mathcal{J}} \xi_j \left(\text{tr}\left(\mathbf{G}_j\mathbf{V}\right) + \sum_{i\in\mathcal{I}^u} \text{tr}\left(\mathbf{G}_j\mathbf{W}_i\right)\right) - \mathbf{\Gamma}\mathbf{V} \\ &- \sum_{i\in\mathcal{I}^c} \mathbf{\Omega}_i\mathbf{W}_i - \sum_{i\in\mathcal{I}^c} \zeta_i \log(1+\text{tr}\left(\mathbf{H}_i\mathbf{W}_i\right)/\sigma_i^2) \\ &- \sum_{i\in\mathcal{I}^c} \phi_i \text{tr}\left(\mathbf{H}_i\mathbf{W}_i\right)/\sigma_i^2,\end{aligned}$$

where $\mu_i \geq 0$, $\xi_j \geq 0$, $\mathbf{\Gamma} \succeq \mathbf{0}$, $\mathbf{\Omega}_i \succeq \mathbf{0}$, $\zeta_i \geq 0$, $\phi_i \geq 0$ are Lagrange multipliers for constraints (15), (17), (18), (19), (22), and (23) respectively. $\mathbf{\Gamma}$ and $\mathbf{\Omega}_i$ are both $JK \times JK$ matrices. Then

$$\begin{aligned}\frac{\partial \mathscr{L}}{\partial \mathbf{W}_i} =& \eta \mathbf{I} - \frac{\zeta_i}{\sigma_i^2 \ln 2}(1 + \mathrm{tr}\,(\mathbf{H}_i \mathbf{W}_i)/\sigma_i^2)^{-1}\mathbf{H}_i - \frac{\phi_i}{\sigma_i^2 \ln 2}\mathbf{H}_i \\ &+ \sum_{j \in \mathcal{J}} \xi_j \mathbf{G}_j - \mathbf{\Omega}_i, \end{aligned} \tag{27}$$

$$\begin{aligned}\frac{\partial \mathscr{L}}{\partial \mathbf{V}} =& \eta \mathbf{I} - \sum_{i \in \mathcal{I}^c} \frac{\mu_i}{\sigma_i^2 \ln 2}(1 + \mathrm{tr}\,(\mathbf{H}_i \mathbf{V})/\sigma_i^2)^{-1}\mathbf{H}_i \\ &+ \sum_{j \in \mathcal{J}} \xi_j \mathbf{G}_j - \mathbf{\Gamma}, \end{aligned} \tag{28}$$

where $\mathbf{I}$ is a $JK \times JK$ identity matrix.

Since $\mathbf{W}_i^*$ is the optimal solution, we have

$$\frac{\partial \mathscr{L}}{\partial \mathbf{W}_i^*} = \mathbf{0}, \text{ and} \tag{29}$$

$$\mathbf{\Omega}_i \mathbf{W}_i^* = \mathbf{0}. \tag{30}$$

Combing (27) and (29), which yields

$$\begin{aligned}\mathbf{\Omega}_i =& \eta \mathbf{I} + \sum_{j \in \mathcal{J}} \xi_j \mathbf{G}_j \\ &- \frac{1}{\sigma_i^2 \ln 2}\left(\zeta_i (1 + \mathrm{tr}\,(\mathbf{H}_i \mathbf{W}_i^*)/\sigma_i^2)^{-1} + \phi_i\right)\mathbf{H}_i. \end{aligned} \tag{31}$$

In the right hand side of (31), the first two terms construct a matrix with full rank, i.e., rank $= JK$. In addition, in (31), the coefficient of the last term $\mathbf{H}_i$ is negative. And recalling that $\mathbf{\Omega}_i \succeq \mathbf{0}$, and rank$(\mathbf{H}_i) \leq 1$, we can conclude

$$\mathrm{rank}(\mathbf{\Omega}_i) \geq JK - 1. \tag{32}$$

Further, combing (30) and (32), we can obtain rank$(\mathbf{W}_i^*) \leq 1$.

Similar to (31), we can also get,

$$\mathbf{\Gamma} = \eta \mathbf{I} + \sum_{j \in \mathcal{J}} \xi_j \mathbf{G}_j - \sum_{i \in \mathcal{I}^c} \frac{\mu_i}{\sigma_i^2 \ln 2}(1 + \mathrm{tr}\,(\mathbf{H}_i \mathbf{V}^*)/\sigma_i^2)^{-1}\mathbf{H}_i. \tag{33}$$

However, in the right hand side of (33), the third term is the summation of multiple rank one matrices. Therefore, we cannot claim rank$(\mathbf{\Gamma}_s) \geq JK - 1$, and as a result, we cannot conclude that rank$(\mathbf{V}^*) \leq 1$.

This completes the proof.

References

[1] 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.*, 2019.

[2] 3GPP, "Study on new radio access technology physical layer aspect (release 14)," TR 38.802, Mar. 2017.

[3] M. Shafi, A. F. Molisch, P. J. Smith, T. Haustein, P. Zhu, P. D. Silva, F. Tufvesson, A. Benjebbour, and G. Wunder, "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.

[4] S. A. Hashemi, C. Condo, F. Ercan, and W. J. Gross, "On the performance of polar codes for 5G eMBB control channel," in *Proc. Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, Pacific Grove, CA, USA, Oct. 2017, pp. 1764–1768.

[5] H. Ji, S. Park, J. Yeo, Y. Kim, J. Lee, and B. Shim, "Ultra reliable and low latency communications in 5G downlink: Physical layer aspects," *IEEE Wireless Commun.*, vol. 25, no. 3, pp. 124–130, Jun. 2018.

[6] Z. Dawy, W. Saad, A. Ghosh, J. G. Andrews, and E. Yaacoub, "Toward massive machine type cellular communications," *IEEE Wireless Commun.*, vol. 24, no. 1, pp. 120–128, Feb. 2017.

[7] K. Samdanis, X. Costa-Perez, and V. Sciancalepore, "From network sharing to multi-tenancy: The 5G network slice broker," *IEEE Commun. Mag.*, vol. 54, no. 7, pp. 32–39, Jul. 2016.

[8] X. Foukas, G. Patounas, A. Elmokashfi, and M. K. Marina, "Network slicing in 5G: Survey and challenges," *IEEE Commun. Mag.*, vol. 55, no. 5, pp. 94–100, May 2017.

[9] NGMN, "Description of network slicing concept," Version 1.0, Jan. 2016.

[10] J. Tang, R. Wen, T. Q. S. Quek, and M. Peng, "Fully exploiting cloud computing to achieve a green and flexible C-RAN," *IEEE Commun. Mag.*, vol. 55, no. 11, pp. 40–46, Nov. 2017.

[11] M. Tao, E. Chen, H. Zhou, and W. Yu, "Content-centric sparse multicast beamforming for cache-enabled cloud RAN," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6118–6131, Sep. 2016.

[12] J. Tang and T. Q. S. Quek, "The role of cloud computing in content-centric mobile networking," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 52–59, Aug. 2016.

[13] R. Wen, J. Tang, T. Q. S. Quek, G. Feng, G. Wang, and W. Tan, "Robust network slicing in software-defined 5G networks," in *Proc. IEEE GLOBECOM*, Singapore, Dec. 2017, pp. 1–6.

[14] N. Zhang, Y. F. Liu, H. Farmanbar, T.-H. Chang, M. Hong, and Z. Q. Luo, "Network slicing for service-oriented networks under resource constraints," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 11, pp. 2512–2521, Nov. 2017.

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

[16] S. Xu, T. H. Chang, S. C. Lin, C. Shen, and G. Zhu, "Energy-efficient packet scheduling with finite blocklength codes: Convexity analysis and efficient algorithms," *IEEE Trans. Wireless Commun.*, vol. 15, no. 8, pp. 5527–5540, Aug. 2016.

[17] K. I. Pedersen, G. Berardinelli, F. Frederiksen, P. Mogensen, and A. Szufarska, "A flexible 5G frame structure design for frequency-division duplex cases," *IEEE Commun. Mag.*, vol. 54, no. 3, pp. 53–59, Mar. 2016.

[18] G. Pocovi, K. I. Pedersen, B. Soret, M. Lauridsen, and P. Mogensen, "On the impact of multi-user traffic dynamics on low latency communications," in *Proc. IEEE ISWCS*, Poznan, Poland, Sep. 2016, pp. 204–208.

[19] C. Lu and Y. F. Liu, "An efficient global algorithm for single-group multicast beamforming," *IEEE Trans. Signal Process.*, vol. 65, no. 14, pp. 3761–3774, Jul. 2017.

[20] H. T. Wai, Q. Li, and W.-K. Ma, "Discrete sum rate maximization for MISO interference broadcast channels: Convex approximations and efficient algorithms," *IEEE Trans. Signal Process.*, vol. 64, no. 16, pp. 4323–4336, Aug. 2016.

[21] Z.-Q. Luo, W.-K. Ma, A. M.-C. So, Y. Ye, and S. Zhang, "Semidefinite relaxation of quadratic optimization problems," *IEEE Signal Process. Mag.*, vol. 27, no. 3, pp. 20–34, May 2010.

[22] S. Schiessl, J. Gross, and H. Al-Zubaidy, "Delay analysis for wireless fading channels with finite blocklength channel coding," in *Proc. ACM MSWiM*, Cancun, Mexico, Nov. 2015, pp. 13–22.

[23] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," http://cvxr.com/cvx, Mar. 2014.

[24] X. Zheng, X. Sun, D. Li, and J. Sun, "Successive convex approximations to cardinality-constrained convex programs: a piecewise-linear DC approach," *Computational Optimization and Applications*, vol. 59, no. 1, pp. 379–397, 2014.

[25] M. Razaviyayn, H.-W. Tseng, and Z.-Q. Luo, "Computational intractability of dictionary learning for sparse representation," 2015. [Online]. Available: http://arxiv.org/abs/1511.01776

[26] J. Tang, T. Q. S. Quek, T.-H. Chang, and B. Shim, "Systematic resource allocation in cloud RAN with caching as a service under two timescales." [Online]. Available: https://www.dropbox.com/s/s447g0zd0404wl6/

[27] 3GPP, "LTE; Evolved universal terrestrial radio access (E-UTRA); Radio frequency (RF) requirements for LTE Pico Node B (release 9)," 3rd Generation Partnership Project (3GPP), TS 36.931, May 2011, v9.0.0.