Traffic Steering for eMBB and uRLLC Coexistence in Open Radio Access Networks

Fatemeh Kavehmadavani, Van-Dinh Nguyen, Thang X. Vu, and Symeon Chatzinotas
Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg
Email: {fatemeh.kavehmadavani, dinh.nguyen, thang.vu, symeon.chatzinotas}@uni.lu

***Abstract*—Existing radio access network (RAN) architectures are lack of sufficient openness, flexibility, and intelligence to meet the diverse demands of emerging services in beyond 5G and 6G wireless networks, including enhanced mobile broadband (eMBB) and ultra-reliable and low-latency (uRLLC). Open RAN (ORAN) is a promising paradigm that allows building a virtualized and intelligent architecture. In this paper, we focus on traffic steering (TS) scheme based on multi-connectivity (MC) and network slicing (NS) techniques to efficiently allocate heterogeneous network resources in "NextG" cellular networks. We formulate the RAN resource allocation problem to simultaneously maximize the weighted sum eMBB throughput and minimize the worst-user uRLLC latency subject to QoS requirements, and orthogonality, power, and limited fronthaul constraints. Since the formulated problem is categorized as a mixed integer nonlinear problem (MINLP), we first relax binary variables to continuous ones and develop an efficient iterative algorithm based on successive convex approximation technique. System-level simulation results demonstrate the effectiveness of the proposed algorithm, compared to several well-known benchmark schemes.**

I. Introduction

Beyond fifth-generation (5G) and sixth-generation (6G) wireless networks are expected to meet three major services with different demands, *namely* ultra Reliability Low-Latency Communication (uRLLC), enhanced Mobile Broadband (eMBB), and massive Machine-Type Communications (mMTC). Different uRLLC, which supports traffics requiring extremely high reliability (i.e. 99.999%) and very low latency (i.e. less than 1 ms), eMBB requires high throughput connectivity [1]. Since existing 5G cellular networks are inflexible, closed, and aggregated, it is not possible to enable the coexistence of various services on the existing "one-size-fits-all" 5G architecture. Despite the cost-effective of centralized radio access network (CRAN) and virtual RAN (vRAN), these architectures are still lack of open interfaces, non-proprietary hardware and software. Therefore, open RAN (ORAN) has been recently proposed to address these issues, evolving towards a flexible, virtualized, disaggregated, open, and intelligent "NextG" wireless networks [2].

The key objective of ORAN is to improve the RAN performance by virtualizing RAN's elements, disaggregation of their components, software, and hardware, defining appropriate open interfaces for connecting them, as well as embedding machine learning (ML)/artificial intelligence (AI) techniques to construct adaptive and smarter RAN layers in its architecture [3]. Two novel modules, including near-real-time RAN intelligent controller (near-RT RIC) and non-RT RIC, are defined to enable a centralized network abstraction to further reduce cost, network complexity, and human-machine interaction [4]. Following a disaggregation approach, base station (BS) functionalities are virtualized as network functions based on 3GPP functional split and are divided across various network nodes, *namely* central unit (CU), distributed unit (DU), and radio unit (RU).

Mobile networks have become increasingly complicated as network capacity and traffic have increased manyfold. It is challenging to steer heterogeneous traffics to effectively improve network efficiency and user experience [5]. Network slicing (NS) has appeared as a promising solution to allocate resources to various wireless services with different requirements [6]. To meet the strict latency requirement of uRLLC services, 5G NR adopts dynamic numerologies for a short transmission duration and also introduces the concept of mini-slots by diminishing the transmission time interval (TTI), which can be achievable with a larger subcarrier spacing [7]. Because of the multi-link feature, multi-connectivity (MC) is the most effective strategy for realizing eMBB and uRLLC coexistence, that improves signal-to-interference-plus-noise ratio (SINR) and increases user reliability and communication coverage.

Traffic steering (TS) is crucial to illustrate the practical applicability of ORAN architecture. However, the current literature on about TS is still sparse and isolated. In [8], a dynamic MC-based joint scheduling framework with TS for both eMBB and uRLLC traffics was proposed. Zhang *et al.* discussed TS in LTE networks with unlicensed bands in [9]. Due to the impact of the NS technique in multi-services networks, the authors in [10] analyzed schemes to enable dynamic TS and energy-efficient RAN moderation in 5G. To the best of our knowledge, there is no work in the literature to model TS in the ORAN architecture.

Unlike typical TS mechanisms that treat all users in the same way without considering user demands, in this paper we propose a novel TS scheme using dynamic MC technique and RAN slicing scheme to steer traffic flows towards the most suitable cells based on user-centric condition. Our main contributions are summarized as follows:

- We develop a joint RAN resource allocation framework as xAPP in near-RT RIC to dynamically optimize eMBB and uRLLC traffics over the same resources. The utility function of interest combines the sum eMBB throughput and worst-user uRLLC latency. In addition, a rigorous

This work was supported in part by the ERC AGNOSTIC project, ref. H2020/ERC2020POC/957570.

analysis for the uRLLC latency model will be provided that takes into account all factors of computation and communication. Existing RUs in the site will generate MC clusters, which vary according to the characteristics of the network. This, in turn, achieves a higher sum eMBB throughput while maintaining the uRLLC latency requirement.

- We propose an efficient iterative algorithm based on successive convex approximation (SCA) method to solve the relaxed optimization problem, which guarantees at least a locally optimal solution.
- Numerical results are provided to show a fast convergence behavior of the proposed algorithm and verify its effectiveness, compared with benchmark schemes.

II. System Model

A. Signal Model

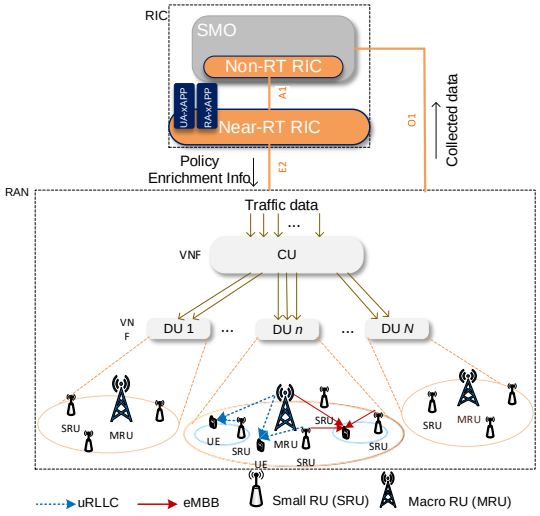

Fig. 1. ORAN-based system model.

We consider a downlink orthogonal frequency-division multiple access (OFDMA) in the ORAN architecture, consisting of one CU and a set $\mathcal{N} \triangleq \{1, 2, \ldots, N\}$ of N DUs. These two types of processing nodes run on the general-purpose data centers as virtual machines (VMs) by virtual network function (VNF) technology, which can process incoming user packets in parallel, as illustrated in Fig. 1. Towards cost-efficient deployment, we assume that DUs serve non-overlapped geographical areas, i.e. each DU serves a cluster of RUs.

Let $\mathcal{M}_n \triangleq \{0, 1, \ldots, m, \ldots, M_n\}$ indicates the set of RUs served by the n-th DU, which consists of one macro RU (MRU), i.e. RU 0, and M_n small-cell RUs (SRUs), i.e. RU $m \in \{1, \cdots, M_n\}$. SRUs' unique ability to handle high-density data makes them the ideal choice to satisfy demands of services in 5G networks, especially uRLLC. In contrast, the MRU provides a high data rate and an extended coverage to eMBB users. SRUs are also used to support the control plane and provide a quick response to small packet data size [11]. In addition, we denote by $\mathcal{U}_n = \{1, \ldots, U_n\}$ the set of users served by DU n, which can be further divided into two disjoint sets $\mathcal{U}_n^{\mathsf{ur}}$ of U_n^{ur} uRLLC users, and $\mathcal{U}_n^{\mathsf{em}}$ of U_n^{em} eMBB users. The m-th RU is equipped with K_m antennas while users are equipped with a single antenna and are randomly distributed across the network area.

Under MC configuration, the MRU's operating frequency is orthogonal to that of SRUs. Denote by $\mathcal{F}_0$ and $\mathcal{F}_1$ the sets of sub-band frequencies operated by MRU and SRUs, respectively. The numbers of sub-bands operated by MRU and SRUs are $F_0 = |\mathcal{F}_0|$ and $F_1 = |\mathcal{F}_1|$, respectively. Each transmission frame is divided into T time slots whose duration is equal to one TTI. The bandwidth of each sub-band defined in 3GPP 5G NR is equal to 360 KHz. Therefore, one resource block (RB) with the time duration of mini-slot ($\Delta_t = 0.25$ ms) corresponds to 7 OFDM symbols.

B. Achievable Rate

Our goal is to optimize the RAN performance in 3D, encompassing time, frequency, and power domains. The problem casts into assigning the total $(F_0 + F_1)T$ RBs of all RUs covered by the considered DU to its users. Because of the non-overlapped DUs' coverage, the resource optimization design at one DU is similar to that of other DUs. Thus, for ease of presentation, we drop the subscript index of DUs hereafter.

Let $\boldsymbol{h}_{t,f,m,u} \in \mathbb{C}^{K_m \times 1}$ be the channel vector from the m-th RU to the u-th UE at the sub-band f and the time slot t, including the path-loss. Within each frame, assume that the channel remains temporally invariant, while it may be different across the sub-bands. In this work, we employ maximal ratio transmission (MRT) to maximize the received SNR, which is also the optimal beamformer under the considered OFDMA scheme [12]. As a result, the effective channel gain for UE u served by RU m at sub-band f and time slot t is $g_{t,f,m,u} \triangleq \|\boldsymbol{h}_{t,f,m,u}\|_2^2$.

Given the orthogonality constraint, this work considers that each RB of a RU is assigned to only one single user during one time slot. To introduce this assignment we define the decision variables $\pi_{t,f,m,u}^{\mathsf{em}} \in \{0, 1\}$ and $\pi_{t,f,m,u}^{\mathsf{ur}} \in \{0, 1\}$ for eMBB and uRLLC traffics, respectively. Here, $\pi_{t,f,m,u}^{\mathsf{em}} = 1$ if the RB associated with time-slot t and sub-band f of RU m assigned to the u-th eMBB user, and $\pi_{t,f,m,u}^{\mathsf{em}} = 0$ otherwise; similarly definition for uRLLC users. The achievable rate in bits/s for a given set of channel realizations at the u-th eMBB user is given by:

$$r_u^{\mathsf{em}}(\boldsymbol{p}^{\mathsf{em}}) = \sum_m \sum_t \sum_f \beta \log_2\Big(1 + \frac{p_{t,f,m,u}^{\mathsf{em}} g_{t,f,m,u}}{N_0}\Big) \quad (1)$$

where β, N_0 and $p_{t,f,m,u}^{\mathsf{em}}$ are the bandwidth of each RB, power of the Additive White Gaussian Noise (AWGN), and transmit power from RU m to UE u for eMBB traffic at sub-band f at the TTI t, respectively. Let us define $\boldsymbol{p}^{\mathsf{em}} \triangleq [p_{t,f,m,u}^{\mathsf{em}}]_{\forall t,f,m,u}$. We note that there is no inter-user interference in (1) since it is cancelled via OFDMA constraints introduced latter in this section. Furthermore, the transmit power must satisfy $p_{t,f,m,u}^{\mathsf{em}} \leq \pi_{t,f,m,u}^{\mathsf{em}} P_m^{\max}$ with $P_m^{\max}$ being the power budget at RU m, which guarantees that RU m allocates power to user u on RB (t, f) only if $\pi_{t,f,m,u}^{\mathsf{em}} = 1$; otherwise $\pi_{t,f,m,u}^{\mathsf{em}} = 0$ and

$p^{\mathsf{em}}_{t,f,m,u} = 0$. As a result, the sum throughput of eMBB users is given as $\mathcal{R}^{\mathsf{em}}(\boldsymbol{p}^{\mathsf{em}}) = \sum_{u\in\mathcal{U}^{\mathsf{em}}} r^{\mathsf{em}}_u(\boldsymbol{p}^{\mathsf{em}})$. The minimum QoS requirement for eMBB users is guaranteed by the constraint $r^{\mathsf{em}}_u(\boldsymbol{p}^{\mathsf{em}}) \geq R_{th}$, where R_{th} is a given QoS threshold.

In contrast, owing to the finite block-length in uRLLC traffics, the Shannon capacity is no longer used to obtain the throughput for uRLLC users. The achievable rate of u-th user for uRLLC traffic using the short block-length can be expressed as [13]

$$r^{\mathsf{ur}}_u(\boldsymbol{p}^{\mathsf{ur}}, \boldsymbol{\pi}^{\mathsf{ur}}) = \sum\nolimits_{t,f,m} \beta \Big[\log_2\Big(1 + \frac{p^{\mathsf{ur}}_{t,f,m,u} g_{t,f,m,u}}{N_0}\Big) - \frac{\pi^{\mathsf{ur}}_{t,f,m,u}\sqrt{V} Q^{-1}(P_e)}{\sqrt{\Delta_t \beta}}\Big] \quad (2)$$

where V, P_e and Q^{-1}: $\{0,1\} \rightarrow \mathbb{R}$ denote the channel dispersion, error probability, and inverse of the Gaussian Q-function, respectively. It is observed that $V = 1 - \frac{1}{(\mathrm{SNR})^2} \approx 1$ when the received SNR=$\frac{p^{\mathsf{ur}}_{t,f,m,u} g_{t,f,m,u}}{N_0} \geq \Gamma_0$ with $\Gamma_0 \geq 5$ dB. This can be easily achieved in cellular networks, by arranging the uRLLC decoding vector into one possible null space of the reference subspace [14]. Hence, we consider the constraint $\frac{N_0\beta}{g_{t,f,m,u}}\Gamma_0 \pi^{\mathsf{ur}}_{t,f,m,u} \leq p^{\mathsf{ur}}_{t,f,m,u} \leq \pi^{\mathsf{ur}}_{t,f,m,u} P^{max}_m$ to guarantee the approximation $V \approx 1$ as well as the big-M formulation theory to avoid non-convexity of (2).

C. uRLLC Latency Model

Unlike eMBB users, the main aim for uRLLC users is to minimize their end-to-end (e2e) latency from CU to end users. To meet stringent latency requirements, uRLLC traffics should be immediately served, resulting in no queuing delay. As a result, the e2e latency of uRLLC users consists of the processing and transmission time.

As depicted in Fig. 1, after processing, the uRLLC users's arrival packets at the CU layer are subsequently routed to VNFs in the DU layer for parallel processing. We adopt the $M/M/1$ processing queue model on the first-come first-serve basis to serve each user's packets. The packet process of the u-th arrival packet for uRLLC traffics generated by the FTP3 model standardized in 3GPP with Z bytes length follows the Poisson process with the mean arrival rate λ_u (packets/frame). It is assumed that each packet has an identical length, and packet segmentation is not allowed. Suppose that each assigned RB to the uRLLC user should transmit at least one complete data packet. In contrast, eMBB users generate continuous traffic with infinite packet size, avoiding packet loss due to buffer overflow.

Denote by f_{cu} and f_{du} the computation capacities of CU and DU (cycles/sec), respectively. Considering the identical packet size, the required computation resource to process one packet of size Z is C (number of cycles). As result, $\mu_{cu} = f_{cu}/C$ and $\mu_{du} = f_{du}/C$ are the task rates (1/sec) at CU and DU, respectively. As a result, $1/\mu_{cu}$ and $1/\mu_{du}$ represent the mean service time of CU and DU layers, respectively. The processing latency of the uRLLC arrival packet at the CU layer (τ^{pro}_{cu}) and DU layer (τ^{pro}_{du}) is computed as

$$\tau^{\mathsf{pro}}_{cu} = 1/(\mu_{cu} - \Lambda) \text{ and } \tau^{\mathsf{pro}}_{du} = 1/(\mu_{du} - \Lambda) \quad (3)$$

where $\Lambda = \sum_{u\in\mathcal{U}^{\mathsf{ur}}} \lambda_u$ is the total packet arrival rate for all uRLLC users at the CU layer. We assume that $\mu_{cu} > \Lambda$ and $\mu_{du} > \Lambda$ to guarantee the queue stability.

Next, the arrival packets λ_u for the u-th uRLLC user is transported to DU via the midhauld (MH) link with the maximum capacity C^{MH} (bits/sec). It should be mentioned that the mean arrival data rate of the DU layer is approximately equal to the mean arrival data rate of the first layer. By the Burke's theorem, the mean arrival data rate of the second layer, which is processed in the first layer, is still Poisson with the rate Λ [15]. Hence, the data transmission latency of the uRLLC traffic for UE u under the MH limited capacity is:

$$\tau^{\mathsf{tx}}_{cu,du} = \frac{\Lambda Z}{C^{\mathsf{MH}}}. \quad (4)$$

Using the MC technique, the generated traffic per frame for the u-th uRLLC user is split into several partitions which are transmitted in separate links, and then aggregated at this user. The maximum number of paths from DU n to each user is M_n. We denote by $\boldsymbol{b}_u \triangleq [b_{0,u}, b_{1,u}, \cdots, b_{M_n,u}]$ the flow-split indicator vector for the u-th uRLLC user. In particular, if $b_{m,u} = 1$, RU $m \in \mathcal{M}_n$ is selected to transmit the data of the u-th uRLLC user; otherwise $b_{m,u} = 0$. In addition, let us denote by $\boldsymbol{\varphi}_u \triangleq [\varphi_{0,u}, \varphi_{1,u}, \varphi_{2,u}, \cdots, \varphi_{M_n,u}]$ the flow-split portion vector of user u with $\sum_{m\in\mathcal{M}_n} \varphi_{m,u} = 1$, where $\varphi_{m,u}$ represents a portion of traffic routed to user u via RU m. Since the packets for the u-th user can be transmitted by multiple RUs, the effective response time $\tau^{\mathsf{tx}}_{du,ru}$ to transport all packets the DU layer should be computed by the worst average response time among its connected fronthaul (FH) links with maximum capacity C^{FH}_m (bits/sec), i.e.

$$\tau^{\mathsf{tx}}_{du,ru} = \max_m \Big\{ \frac{\sum_{u\in\mathcal{U}^{\mathsf{ur}}_n} \varphi_{m,u}\lambda_u Z}{C^{\mathsf{FH}}_m} \Big\}, \ \forall m \in \mathcal{M}_n. \quad (5)$$

We denote by $r^{\mathsf{ur}}_{m,u}$ the data rate from RU m to user u, which can be directly extracted from (2). The transmission latency from RU m to user u is then calculated as

$$\tau^{\mathsf{tx}}_{ru,u} = \max_m \Big\{ \frac{\varphi_{m,u}\lambda_u Z}{r^{\mathsf{ur}}_{m,u}} \Big\}, \quad \forall u \in \mathcal{U}^{\mathsf{ur}}. \quad (6)$$

Simply put, the e2e latency of uRLLC user $u \in \mathcal{U}^{\mathsf{ur}}$ per frame is computed as

$$\tau^{\mathsf{ur}}_u = \tau^{\mathsf{pro}}_{cu} + \tau^{\mathsf{tx}}_{cu,du} + \tau^{\mathsf{pro}}_{du} + \tau^{\mathsf{tx}}_{du,ru} + \tau^{\mathsf{tx}}_{ru,u} + \tau^{\mathsf{pro}}_{ru} + \tau^{\mathsf{pro}}_{u} + \tau^{\mathsf{align}} \quad (7)$$

where $\tau^{\mathsf{pro}}_{ru}, \tau^{\mathsf{pro}}_{u}$ and τ^{align} are the process latency at RU m, uRLLC user u and frame alignment time, respectively. τ^{pro}_{ru} and τ^{pro}_{u} are bounded by three OFDM symbols duration that are typically very small; τ^{align} is upper-bounded by one mini-slot TTI interval. To ensure a minimum latency requirement for uRLLC user u, the e2e latency is bound by a predetermined threshold D_u, i.e. $\tau^{\mathsf{ur}}_u \leq D_u$.

III. Joint Traffic Steering and Resource Allocation Optimization

A. Problem Formulation

Utility function: The main goal is to jointly optimize the RAN slicing and TS to serve eMBB and uRLLC users, subject to various resource constraints and diverse QoS requirements. To doing so, the utility function should capture both the

sum eMBB throughput and worst-user e2e uRLLC latency, such as: $\alpha\frac{\mathcal{R}^{\mathsf{em}}}{R_0}-(1-\alpha)\max_u\left\{\frac{\tau_u^{\mathsf{ur}}}{\tau_0}\right\}$, where $R_0>0$ and $\tau_0>0$ are the reference throughput of eMBB and latency of uRLLC, respectively, which are used to balance two different dimensions of the two quantities; and $\alpha\in[0,1]$ denotes the priority parameter. Based on the above definitions and discussions, the problem of joint TS and resource allocation is mathematically formulated as

$$\max_{\boldsymbol{\varphi},\boldsymbol{\pi},\boldsymbol{p}}\quad \alpha\frac{\mathcal{R}^{\mathsf{em}}}{R_0}-(1-\alpha)\max_u\left\{\frac{\tau_u^{\mathsf{ur}}}{\tau_0}\right\} \tag{8a}$$

$$\text{s.t. C1:}\quad \pi^{\mathsf{em}}_{t,f,m,u},\ \pi^{\mathsf{ur}}_{t,f,m,u}\in\{0,1\},\ \forall t,f,m,u \tag{8b}$$

$$\text{C2:}\ \sum_{t,f,m,u}[\pi^{\mathsf{em}}_{t,f,m,u}+\pi^{\mathsf{ur}}_{t,f,m,u}]\le 1,\forall m=0,f\in\mathscr{F}_0$$
$$\sum_{t,f,m,u}[\pi^{\mathsf{em}}_{t,f,m,u}+\pi^{\mathsf{ur}}_{t,f,m,u}]\le 1,\forall m\neq 0,f\in\mathscr{F}_1 \tag{8c}$$

$$\text{C3:}\ \sum\nolimits_{m,t,f}\pi^{\mathsf{ur}}_{t,f,m,u}\ge e^{\mathsf{ur}}_u,\quad \forall u\in\mathscr{U}^{\mathsf{ur}} \tag{8d}$$

$$\text{C4:}\ \sum\nolimits_{f,u}(p^{\mathsf{em}}_{t,f,m,u}+p^{\mathsf{ur}}_{t,f,m,u})\le P^{\max}_m,\ \forall t,m \tag{8e}$$

$$\text{C5:}\ 0\le p^{\mathsf{em}}_{t,f,m,u}\le\pi^{\mathsf{em}}_{t,f,m,u}P^{\max}_m,\ \forall t,f,m,u \tag{8f}$$

$$\text{C6:}\ \frac{N_0\Gamma_0\pi^{\mathsf{ur}}_{t,f,m,u}}{g_{t,f,m,u}}\le p^{\mathsf{ur}}_{t,f,m,u}\le\pi^{\mathsf{ur}}_{t,f,m,u}P^{\max}_m,\quad \forall t,f,m,u \tag{8g}$$

$$\text{C7:}\ r^{\mathsf{em}}_u(\boldsymbol{p}^{\mathsf{em}})\ge R_{th},\quad \forall u\in\mathscr{U}^{\mathsf{em}} \tag{8h}$$

$$\text{C8:}\ \sum\nolimits_u\left[r^{\mathsf{em}}_{m,u}(\boldsymbol{p}^{\mathsf{em}})+r^{\mathsf{ur}}_{m,u}(\boldsymbol{\pi}^{\mathsf{ur}},\boldsymbol{p}^{\mathsf{ur}})\right]\le C^{\mathsf{FH}}_m,\forall m \tag{8i}$$

$$\text{C9:}\quad r^{\mathsf{ur}}_{m,u}(\boldsymbol{\pi}^{\mathsf{ur}},\boldsymbol{p}^{\mathsf{ur}})\ge\frac{\varphi_{m,u}\lambda_{m,u}Z}{T},\ \forall m,u\in\mathscr{U}^{\mathsf{ur}} \tag{8j}$$

$$\text{C10:}\ \sum\nolimits_m\varphi_{m,u}=1,\ 0\le\varphi_{m,u}\le 1,\ \forall u\in\mathscr{U}^{\mathsf{ur}} \tag{8k}$$

$$\text{C11:}\quad \tau^{\mathsf{ur}}_u(\boldsymbol{\varphi}_u,\boldsymbol{\pi}^{\mathsf{ur}},\boldsymbol{p}^{\mathsf{ur}})\le D_u,\ \forall u\in\mathscr{U}^{\mathsf{ur}} \tag{8l}$$

where $\boldsymbol{\varphi},\boldsymbol{\pi}$ and $\boldsymbol{p}$ are the vectors, encompassing the flow-split portions, sub-band assignments, and power allocation variables, respectively. Here, C2 is the orthogonality constraint to assure that each RB is allocated to only one user (either eMBB or uRLLC). Constraint C3 refers to QoS requirements for uRLLC users, which expresses that every scheduled uRLLC user should be assigned at least $e^{\mathsf{ur}}_u=\lceil\frac{\lambda Z}{B}\rceil$ number of RBs from the dedicated uRLLC slice to empty the available packets in the queues of uRLLC users; and B is the number of bits that each RB carries. Constraint C4 ensures that the total transmission power is no larger than the power budget at RU m, denoted by $P^{\max}_m$. The limited capacity of FH link between DU and RU m is expressed by constraint C8. Finally, constraint C9 ensures that each RB assigned to the u-th uRLLC user should transmit a complete data packet with size Z.

B. Proposed SCA-based Iterative Algorithm for Solving (8)

Challenges of solving problem (8): The main challenges in solving problem (8) lies in the non-convexity of τ^{ur}_u (appears in the objective function and constraint C11) and constraint C8 with respect to TS and transmit power variables. Furthermore, the binary nature of the sub-band allocation variables makes the problem more difficult to solve it directly. Once may employ the MILP solver, e.g. Gurobi or MOSEK, to directly solve binary $\boldsymbol{\pi}$. However, we argue that the exponential computation complexity of such MIL formulation limits its practical feasibility, especially when the number of variables exceeds few hundreds in ORAN scenarios. To tackle these difficulties, we first relax binary variables to continuous ones (i.e. the box constraints between 0 and 1), and transform constraint τ^{ur}_u and C8 into a more traceable form which can be efficiently solved by SCA-based iterative algorithm.

Penalty function: In order to speed up the convergence of the proposed iterative algorithm presented shortly, we introduce the following penalty function

$$\mathcal{P}(\boldsymbol{\pi})=\sum_{t,f,m,u}\left[(\pi^{\mathsf{em}}_{t,f,m,u})^2+(\pi^{\mathsf{ur}}_{t,f,m,u})^2-\pi^{\mathsf{em}}_{t,f,m,u}-\pi^{\mathsf{ur}}_{t,f,m,u}\right]$$

which is convex in $\boldsymbol{\pi}$. It is clear that $\mathcal{P}(\boldsymbol{\pi})\le 0$ for any $\pi_{t,f,m,u}\in[0,1]$, which is useful to penalize the relaxed variables to obtain near-exact binary solutions at optimum (i.e. satisfying C1). By incorporating $\mathcal{P}(\boldsymbol{\pi})$ into the objective function of (8), the parameterized relaxed problem is expressed as

$$\max_{\boldsymbol{\varphi},\boldsymbol{\pi},\boldsymbol{p}}\quad \alpha\frac{\mathcal{R}^{\mathsf{em}}}{R_0}-(1-\alpha)\max_u\left\{\frac{\tau_u^{\mathsf{ur}}}{\tau_0}\right\}+\gamma\mathcal{P}(\boldsymbol{\pi}) \tag{9a}$$

$$\text{s.t. }\widetilde{\text{C1}}:\ \pi^{\mathsf{em}}_{t,f,m,u},\ \pi^{\mathsf{ur}}_{t,f,m,u}\in[0,1],\ \forall t,f,m,u \tag{9b}$$

$$\text{C2 - C11} \tag{9c}$$

where $\gamma>0$ is a given penalty parameter.

Proposition 1: With an appropriate positive value of γ, problems (8) *and* (9) *share the same optimal solution, i.e.* $(\boldsymbol{\varphi}^*,\boldsymbol{\pi}^*,\boldsymbol{p}^*)$.

The proof is directly followed [16] by showing that $\mathcal{P}(\boldsymbol{\pi})=0$ at optimum in maximizing (9a). It implies that there always exists a constant γ to ensure that $\boldsymbol{\pi}$ are binary at optimum, so that the relaxation is tight. In practice, it is acceptable if $\mathcal{P}(\boldsymbol{\pi})\le\varepsilon$ for a very small ε, which leads to a near-exact optimal solution.

To handle the non-convexity of τ^{ur}_u, we introduce new variables $\boldsymbol{t}\triangleq[t_1,t_2]$ to rewrite (9) equivalently as

$$\max_{\boldsymbol{\varphi},\boldsymbol{\pi},\boldsymbol{p},\boldsymbol{t}}\quad \alpha\frac{\mathcal{R}^{\mathsf{em}}}{R_0}-(1-\alpha)\max_u\left\{\frac{\widetilde{\tau}_u^{\mathsf{ur}}}{\tau_0}\right\}+\gamma\mathcal{P}(\boldsymbol{\pi}) \tag{10a}$$

$$\text{s.t. :}\quad \widetilde{\text{C1}},\ \text{C2 - C10} \tag{10b}$$

$$\widetilde{\text{C11}}:\quad \widetilde{\tau}^{\mathsf{ur}}_u\le D_u,\ \forall u \tag{10c}$$

$$\text{C12:}\quad r^{\mathsf{ur}}_{m,u}\ge 1/t_1,\ \forall m,u \tag{10d}$$

$$\text{C13:}\quad \varphi_{m,u}t_1\le\frac{t_2}{\lambda_u Z},\forall m,u \tag{10e}$$

where $\widetilde{\tau}^{\mathsf{ur}}_u=\tau^{\mathsf{pro}}_{cu}+\tau^{\mathsf{tx}}_{cu,du}+\tau^{\mathsf{pro}}_{du}+\tau^{\mathsf{tx}}_{du,ru}+t_2+\tau^{\mathsf{pro}}_{ru}+\tau^{\mathsf{pro}}_u+\tau^{\mathsf{align}}$. The equivalence between (9) and (10) is easily verified by showing equality of C12 and C13 at optimum. In problem (10), the objective function is non-concave due to $\mathcal{P}(\boldsymbol{\pi})$, while constraints C8 and C13 are non-convex.

Under SCA method, the function $\mathcal{P}(\boldsymbol{\pi})$ is linearized at the κ-th iteration by the first-order Taylor approximation as

$$\mathcal{P}^{(\kappa)}(\boldsymbol{\pi})\triangleq\sum_{t,f,m,u}\left[\pi^{\mathsf{em}}_{t,f,m,u}(2\pi^{\mathsf{em},(\kappa)}_{t,f,m,u}-1)-(\pi^{\mathsf{em},(\kappa)}_{t,f,m,u})^2\right.$$
$$\left.+\pi^{\mathsf{ur}}_{t,f,m,u}(2\pi^{\mathsf{ur},(\kappa)}_{t,f,m,u}-1)-(\pi^{\mathsf{ur},(\kappa)}_{t,f,m,u})^2\right] \tag{11}$$

where $\mathcal{P}(\boldsymbol{\pi}) \geq \mathcal{P}^{(\kappa)}(\boldsymbol{\pi})$ and $\mathcal{P}(\boldsymbol{\pi}^{(\kappa)}) \geq \mathcal{P}^{(\kappa)}(\boldsymbol{\pi}^{(\kappa)})$. For C8, we denote its LHS as $r_m(\boldsymbol{p}) \triangleq \sum_u \left[r^{\mathsf{em}}_{m,u}(\boldsymbol{p}^{\mathsf{em}}) + r^{\mathsf{ur}}_{m,u}(\boldsymbol{p}^{\mathsf{ur}})\right]$, which is concave in $\boldsymbol{p}$ and can be approximated at $\boldsymbol{p}^{(\kappa)}$ as

$$r^{(\kappa)}_m(\boldsymbol{p}) \triangleq r_m(\boldsymbol{p}^{(\kappa)}) - \sum_{t,f,u} \beta \frac{\pi^{\mathsf{ur}}_{t,f,m,u} Q^{-1}(P_e)}{\sqrt{\Delta_t \beta}} + \frac{\beta}{\ln 2} \sum_{t,f,u,\mathsf{x}} (p^{\mathsf{x}}_{t,f,m,u} - p^{\mathsf{x},(\kappa)}_{t,f,m,u}) \left[\frac{g_{t,f,m,u}}{N_0 + p^{\mathsf{x},(\kappa)}_{t,f,m,u} g_{t,f,m,u}}\right] \quad (12)$$

where $\mathsf{x} \in \{\mathsf{em}, \mathsf{ur}\}$. For C13, it is rewritten equivalently as $(\varphi_{m,u}+t_1)^2 \leq 2\frac{t_2}{\lambda_u Z} + \varphi^2_{m,u} + t_1^2$, where both sides are convex. By the first-order Taylor approximation, we convexify C13 as

$$\widetilde{\mathrm{C}}13: (\varphi_{m,u} + t_1)^2 \leq \frac{2t_2}{\lambda_u Z} + \Phi^{(\kappa)}(\varphi_{m,u}, t_1) \quad (13)$$

where $\Phi^{(\kappa)}(\varphi_{m,u}, t_1) \triangleq 2\varphi^{(\kappa)}_{m,u}\varphi_{m,u} + 2t^{(\kappa)}_1 t_1 - (\varphi^{(\kappa)}_{m,u})^2 - (t^{(\kappa)}_1)^2$.

Bearing all the above approximations in mind, the convex approximate program of (10) solved at iteration κ is given as

$$\max_{\boldsymbol{\varphi},\boldsymbol{\pi},\boldsymbol{p},\boldsymbol{t}} \psi^{(\kappa)} \triangleq \alpha \frac{\mathcal{R}^{\mathsf{em}}}{R_0} - (1-\alpha) \max_u \left\{\frac{\widetilde{\tau}^{\mathsf{ur}}_u}{\tau_0}\right\} + \gamma \mathcal{P}^{(\kappa)}(\boldsymbol{\pi}) \quad (14a)$$

$$\text{s.t.}: \quad \widetilde{\mathrm{C}}1,\ \mathrm{C}2 \text{ - } \mathrm{C}7,\ \mathrm{C}9,\ \mathrm{C}10, \widetilde{\mathrm{C}}11, \mathrm{C}12,\ \widetilde{\mathrm{C}}13 \quad (14b)$$

$$\widetilde{\mathrm{C}}8: \quad r^{(\kappa)}_m(\boldsymbol{p}) \leq C^{\mathsf{FH}}_m, \ \forall m. \quad (14c)$$

The SCA-based iterative algorithm is summarized in Algorithm 1. To guarantee a feasible solution to problem (8), Step 6 is performed to recover an exact binary solution and re-run Steps 1-5 to refine the final solution.

Algorithm 1 SCA-based Iterative Algorithm to Solve (8)

Initialization: Set $\kappa := 1$ and generate initial feasible points for $(\boldsymbol{\varphi}^{(0)}, \boldsymbol{\pi}^{(0)}, \boldsymbol{p}^{(0)}, \boldsymbol{t}^{(0)})$ to constraints in (14)

1: **repeat**
2: Solve (14) to obtain $(\boldsymbol{\varphi}^*, \boldsymbol{\pi}^*, \boldsymbol{p}^*, \boldsymbol{t}^*)$ and ψ^*;
3: Update $(\boldsymbol{\varphi}^{(\kappa)}, \boldsymbol{\pi}^{(\kappa)}, \boldsymbol{p}^{(\kappa)}, \boldsymbol{t}^{(\kappa)}) := (\boldsymbol{\varphi}^*, \boldsymbol{\pi}^*, \boldsymbol{p}^*, \boldsymbol{t}^*)$ and $\psi^{(\kappa)} := \psi^*$;
4: Set $\kappa := \kappa + 1$;
5: **until** Convergence or $|\psi^{(\kappa)} - \psi^{(\kappa-1)}| \leq \delta$ {*/Satisfying a given accuracy level*/}
6: Recover an exact binary by computing $\boldsymbol{\pi}^* = \lfloor \boldsymbol{\pi}^{(\kappa)} + 0.5 \rfloor$ and repeat step 1 to 5 for given $\boldsymbol{\pi}^*$;
7: **Output:** $(\boldsymbol{\varphi}^*, \boldsymbol{\pi}^*, \boldsymbol{p}^*, \boldsymbol{t}^*)$.

Convergence and complexity analysis: Algorithm 1 produces a sequence of improved solutions $\{\boldsymbol{\varphi}^{(\kappa)}, \boldsymbol{\pi}^{(\kappa)}, \boldsymbol{p}^{(\kappa)}, \boldsymbol{t}^{(\kappa)}\}$ (see [17] for more details) as well as a non-decreasing sequence of the objective values $\{\psi^{(\kappa)}\}$, i.e. $\psi^{(\kappa)} \geq \psi^{(\kappa-1)}$. The sequence $\{\psi^{(\kappa)}\}_{\kappa\to\infty}$ is bounded above due to the limited bandwidth and power budgets. By the interior-point method, the per-iteration complexity of Algorithm 1 is $\mathcal{O}\big(\sqrt{c}(v)^3\big)$, where $c = (M_n+1)\big(3U_n(F_0+F_1)T + T + U^{\mathsf{ur}}_n + 1\big) + 3U^{\mathsf{ur}}_n + U^{\mathsf{em}}_n$ and $v = (M_n+1)\big(2U_n(F_0+F_1)T + U^{\mathsf{ur}}_n\big) + 2$ are the numbers of constraints and variables, respectively.

IV. Numerical Results

In this section, we numerically evaluate the performance of the proposed algorithm. All users are uniform randomly located in a circular area with a radius of 500 m. MRU is located at the centered area, serving three sectors, each of which includes two SRUs. The coverage of SRUs is 100 m. The channels are generated as Rayleigh fading with the path-loss $\mathsf{PL}_{\mathrm{MRU-UE}} = 128.1 + 37.6\log_{10}(d/1000)$ for the MRU-UE channels and $\mathsf{PL}_{\mathrm{SRU-UE}} = 38 + 30\log_{10}(d)$ for the SRU-UE channels. The number of frequency sub-bands for MRU and SRUs are $F_0 = 6$ and $F1 = 10$, respectively. Unless stated otherwise, other simulation parameters are given in Table I. For performance comparison, we consider the following two well-known benchmark schemes:

- **Scheme 1**: This scheme optimizes only the traffic steering $\boldsymbol{\varphi}$ and RBs allocation $\boldsymbol{\pi}$, and allocates an equal power to all users.
- **Scheme 2**: The second scheme randomly allocates RBs to users and optimizes only the traffic steering $\boldsymbol{\varphi}$ and power allocation $\boldsymbol{p}$ variables.

Fig. 2 shows the impact of the number of SRUs on the sum eMBB throughput and worst-user uRLLC latency. As expected, the sum eMBB throughput increases and the uRLLC latency decreases as the number of SRUs increases. The sum eMBB throughput achieved by the proposed method is significantly higher than that of benchmark schemes. In addition, the proposed method clearly provides the lowest latency of uRLLC users comparing two other schemes, regardless of the number of SRUs. For instance, the proposed method improves about 50% of the sum eMBB throughput and reduces the uRLLC latency by 90% at $M = 4$, compared to Scheme 1

TABLE I
Simulation Parameters

Parameter	Value	Parameter	Value
No. eMBB users	8	Noise power (N_0)	-110 dBm
No. uRLLC users	4	uRLLC packet size (Z)	32 bytes
BW of MRU	20 MHz	Length of time-frame	10 ms
BW of SRU	100 MHz	Time slot (Δ_t)	0.25 ms
BW of subcarriers (β)	360 KHz	Predetermined latency (D_u)	0.5 ms
Power of MRU	46 dBm	Error probability (P_e)	10^{-3}
Power of SRU	30 dBm	MH capacity	50 Gbps
MRU's FH capacity	1000 Mbps	SRUs' FH capacity	500 Mbps

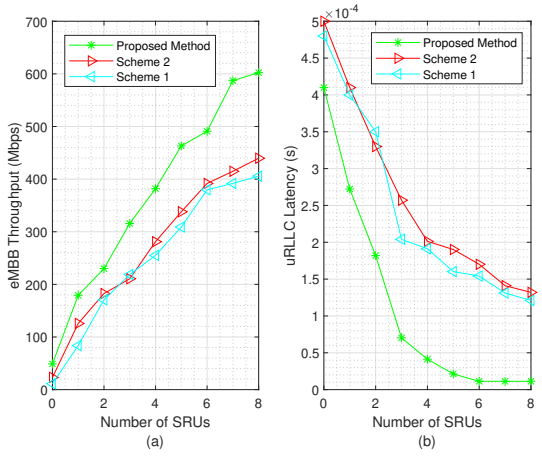

Fig. 2. The sum eMBB throughput and worst-user uRLLC latency versus the number of SRUs.

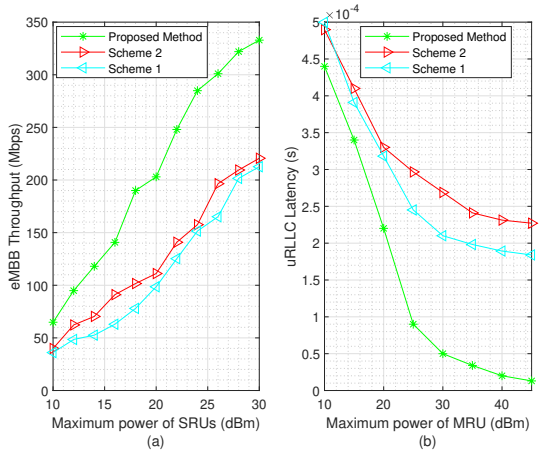

Fig. 3. The sum eMBB throughput and worst-user uRLLC latency versus $P^{\max}$.

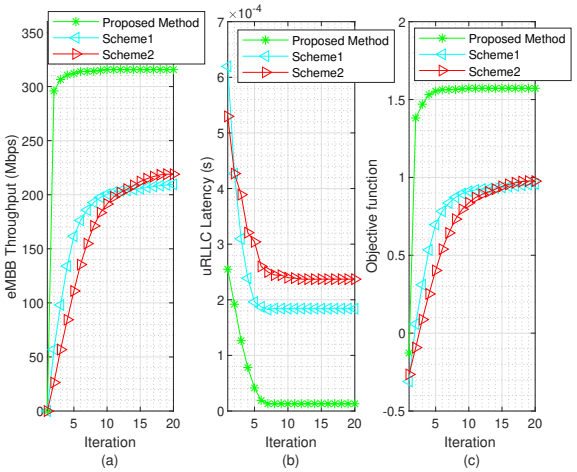

Fig. 4. Convergence behavior of three schemes.

and Scheme 2. These observations confirm the effectiveness of the proposed joint RAN resource allocation framework.

As mentioned previously, the unique characteristics of MRU and SRUs are to meet uRLLC and eMBB demands, respectively. Therefore, we evaluate the impact of the maximum power budgets of these RUs on two services in Fig. 3. Fig. 3(a) plots the sum eMBB throughput as a function of the SRUs' transmit power, while Fig. 3(b) shows the uRLLC latency under the different maximum power of MRU. We can see in both scenarios that, increasing the maximum power of RUs results in an improved sum eMBB throughput and reduced uRLLC latency, which reveal the impact of $P_m^{\max}$ on both objective functions. In addition, according to the performance gain given in Fig. 3, the proposed method offers the highest performance, compared to two considered benchmark schemes. At $P_m^{\max} = 20$ dBm of SRUs (i.e., $\forall m = 1, \cdots, M$), the proposed method achieves 98% and 67% performance gains in terms of the sum eMBB throughput, compered to Scheme 1 and Scheme 2, respectively. Similarly, for $P_0^{\max} = 30$ dBm, the uRLLC latency of the proposed method provides about three times less than Scheme 1 and Scheme 2, respectively.

Finally, we examine the convergence performance of the three considered schemes using Algorithm 1 in Fig. 4. It is shown that the proposed algorithm not only converges very fast reaching the optimal value in less than 10 iterations, but also greatly outperforms two benchmark schemes.

V. Conclusion

We have presented a joint RAN resource allocation framework to realize eMBB and uRLLC coexisting in OFDMA-based ORAN system. We have proposed a comprehensive optimization problem under some practical constraints to maximize the sum eMBB throughput while minimizing the uRLLC latency. We have conducted an in-depth analytical e2e uRLLC latency. A new SCA-iterative algorithm has been developed to solve the formulated problem effectively. We have shown that the proposed method based on MC greatly improves resource utilization, compared to benchmark schemes.

References

[1] P. Popovski *et al.*, "5G wireless network slicing for eMBB, URLLC, and mMTC: A communication-theoretic view," *IEEE Access*, vol. 6, pp. 55765–55779, 2018.

[2] L. Gavrilovska, V. Rakovic, and D. Denkovski, "From cloud RAN to open RAN," *Wirel. Pers. Commun.*, vol. 113, no. 3, pp. 1523–1539, 2020.

[3] L. Bonati *et al.*, "Intelligence and learning in O-RAN for data-driven NextG cellular networks," *IEEE Commun. Mag.*, vol. 59, no. 10, pp. 21–27, 2021.

[4] S. Niknam *et al.*, "Intelligent O-RAN for beyond 5G and 6G wireless networks," 2020. [Online]:. https://arxiv.org/abs/2005.08374.

[5] M. Dryjanski and M. Szydelko, "A unified traffic steering framework for LTE radio access network coordination," *IEEE Commun. Mag.*, vol. 54, no. 7, pp. 84–92, 2016.

[6] S. Vassilaras *et al.*, "The algorithmic aspects of network slicing," *IEEE Commun. Mag.*, vol. 55, no. 8, pp. 112–119, 2017.

[7] M. Iwabuchi *et al.*, "5G field experimental trials on URLLC using new frame structure," in *IEEE Globecom Workshops (GC Wkshps)*, pp. 1–6, 2017.

[8] M. Simsek *et al.*, "Multiconnectivity in multicellular, multiuser systems: A matching-based approach," *Proc. IEEE*, vol. 107, no. 2, pp. 394–413, 2019.

[9] N. Zhang, S. Zhang, S. Wu, J. Ren, J. W. Mark, and X. Shen, "Beyond coexistence: Traffic steering in LTE networks with unlicensed bands," *IEEE Wireless Commun.*, vol. 23, no. 6, pp. 40–46, 2016.

[10] A. Prasad, F. S. Moya, M. Ericson, R. Fantini, and O. Bulakci, "Enabling RAN moderation and dynamic traffic steering in 5G," in *IEEE 84th Veh. Tech. Conf. (VTC-Fall)*, pp. 1–6, 2016.

[11] K. Zhang *et al.*, "Dynamic multiconnectivity based joint scheduling of eMBB and uRLLC in 5G networks," *IEEE Systems Journal*, vol. 15, no. 1, pp. 1333–1343, 2020.

[12] J. Choi and R. W. Heath, "Interpolation based transmit beamforming for mimo-ofdm with limited feedback," *IEEE Transactions on Signal Processing*, vol. 53, no. 11, pp. 4125–4135, 2005.

[13] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2307–2359, 2010.

[14] S. Schiessl *et al.*, "Delay analysis for wireless fading channels with finite blocklength channel coding," in *Proc. 18th ACM Inter. Conf. Model. Anal. and Simul. Wire. and Mob. Sys.*, pp. 13–22, 2015.

[15] P. J. Burke, "The output of a queuing system," *Oper. Res.*, vol. 4, no. 6, pp. 699–704, 1956.

[16] E. Che *et al.*, "Joint optimization of cooperative beamforming and relay assignment in multi-user wireless relay networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 10, p. 5481–5495, 2014.

[17] A. Beck, A. Ben-Tal, and L. Tetruashvili, "A sequential parametric convex approximation method with applications to nonconvex truss topology design problems," *J. Global Optim.*, vol. 47, pp. 29–51, May 2010.