Abstract—The radio access network (RAN) is regarded as one of the potential proposals for massive Internet of Things (mIoT), where the random access channel (RACH) procedure should be exploited for IoT devices to access to the RAN. However, modelling of the dynamic process of RACH of mIoT devices is challenging. To address this challenge, we first revisit the frame and minislot structure of the RAN. Then, we correlate the RACH request of an IoT device with its queue status and analyze the queue evolution process. Based on the analysis result, we derive the closed-form expression of the RA success probability of the device. Besides, considering the agreement on converging different services onto a shared infrastructure, we further investigate the RAN slicing for mIoT and bursty ultra-reliable and low latency communications (URLLC) service multiplexing. Specifically, we formulate the RAN slicing problem as an optimization one aiming at optimally orchestrating RAN resources for mIoT slices and bursty URLLC slices to maximize the RA success probability and energy-efficiently satisfy bursty URLLC slices’ quality-of-service (QoS) requirements. A slice resource optimization (SRO) algorithm exploiting relaxation and approximation with provable tightness and error bound is then proposed to mitigate the optimization problem.

Index Terms—Massive IoT, random access channel, bursty URLLC, RAN slicing

I. INTRODUCTION

With the explosive growth of the Internet of Things (IoT), massive IoT (mIoT) devices, the number of which is predicted to reach 20.8 billion by 2020, will access to the wireless networks for implementing advanced applications, such as e-health, public safety, smart traffic, virtual navigation/management, remote maintenance and control, and environment monitoring. To address the IoT market, the third-generation partnership project (3GPP) has identified mIoT as one of the three main use cases of 5G and has already initiated several task groups to standardize several solutions including extended coverage GSM (EC-GSM), LTE for machine-type communication (LTE-M), and narrowband IoT (NB-IoT) [1], [2].

For establishing massive connections among the wireless networks and mIoT devices, the investigation of reliable and efficient access mechanisms should be prioritized. In accomplishing the massive connections, when an active IoT device wants to transmit signal in the uplink, it randomly chooses a random access (RA) preamble from an RA preamble pool and transmits it through an RA channel (RACH). If more than one device tries to access a base station (BS) simultaneously, then interference occurs at the RRH. During the past few years, a rich body of works on RA mechanisms has been developed [3]–[11] to mitigate interference and improve the RA success probability or reduce the access delay of an IoT device.

Most of the studies [3]–[11], however, assumed that the whole network resources were reserved for the IoT service and did not investigate the case of the coexistence of IoT service and many other services such as enhanced mobile broadband (eMBB) and ultra-reliable and low latency communications (URLLC). The research of the coexistence of IoT service and other services is essential as future networks are convinced to converge variety of services with different latency, reliability, and throughput requirements onto a shared physical infrastructure rather than deploying individual network solution for each service [12]. What is more, owing to the shared characteristic of network resources, some conclusions obtained in the case of providing sole IoT service may become inapplicable if multiple types of services are required to be supported by the networks.

Network slicing is considered as a promising technology in future networks for providing scalability and flexibility in allocating network resources to various services. Recently, many network slicing frameworks have been developed to provide performance guarantees to IoT or massive machine-type communications (mMTC) service, eMBB service, and URLLC service [13]–[19].

Different from previous works, this paper investigates the mIoT and bursty URLLC service multiplexing via slicing the radio access network (RAN). This study is highly challenging because i) performance requirements of a massive number of IoT devices should be satisfied. Yet, the typical 5G cellular IoT, NB-IoT can admit only 50,000 devices per cell [20]; ii) RAN slicing operation (e.g., creating, activating, and releasing slices) has to be conducted in a timescale of minutes to hours to keep pace with the upper layer slicing. However, the wireless channel generally changes in a timescale of millisecond to seconds. Results of the RAN slicing operation are desired to be achieved based on the time-varying channel. Thus, the RAN slicing should tackle a two timescale issue [21].

These challenges motivate us to investigate the RAN slicing...
for mIoT and bursty URLLC service provision to maximize the utility of mIoT slices and that of bursty URLLC slices. The main contributions of this paper can be summarized as the following:

- We revisit the frame and minislot structure for mIoT transmission to accommodate more RA requests from a massive number of IoT devices.
- We adopt a queueing model to track the IoT packet arrival, accumulate and departure processes and analyze the queue evolution process by employing probability and stochastic geometry theories. Based on the analysis result, we derive the closed-form expression of the RA success probability of a randomly chosen IoT device.
- We define mIoT slice utility and bursty URLLC slice utility and formulate the RAN slicing for mIoT and bursty URLLC service multiplexing as a resource optimization problem. The objective of the optimization problem is to maximize the total mIoT and URLLC slice utilities, subject to limited physical resource constraints. The mitigation of this problem is difficult due to the existence of indeterministic objective function and thorny non-convex constraints and the requirement of tackling a two timescale issue as well.
- To mitigate this thorny optimization problem, we propose a slice resource optimization (SRO) algorithm. In this algorithm, we first exploit a sample average approximate (SAA) technique and an alternating direction method of multipliers (ADMM) to tackle the indeterministic objective function and the two timescale issue. Then, a semidefinite relaxation (SDR) scheme joint with a Taylor expansion scheme are leveraged to approximate the non-convex problem as a convex one. The tightness of the SDR scheme and the error bound of the Taylor expansion are also analyzed.

The remaining of this paper is organized as follows. We review the prior arts in Section II. In Section III, we describe our system model and formulate the service multiplexing problem in Section IV. The problem-mitigating algorithm is presented in Sections V and VI. In Section VII, we give the simulation results and conclude this paper in Section VIII.

II. PRIOR ARTS

Recently, many researches have been conducted to increase RA success probabilities and/or reduce the access delay of mIoT devices. They can be generally classified into two groups: traffic detection and estimation based algorithms and algorithms without traffic detection and estimation.

The fundamental idea of the traffic detection and estimation based algorithms is to design an RA algorithm based on the detected and/or estimated users’ activity and traffic congestion situation and so on. For example, to reduce the access delay, a grant-free non-orthogonal RA system relying on the accurate user activity detection and channel estimation was proposed in [3]. A traffic-aware spatiotemporal model for the contention-based RA analysis is conducted for mIoT networks in [7]. With the spatiotemporal model, a hybrid power ramping and back-off RA scheme was then developed to improve the RA success probability. Besides, an extended pseudo-Bayesian backlog estimation scheme was exploited in [11] to estimate the number of backlogged nodes to attempt access. A versatile access control mechanism was then designed to reduce the access delay based on the estimation results.

For algorithms without detection and estimation, they design RA schemes without detecting users’ activity or estimating the statistical characteristic of traffic. For instance, the work in [3] proposed to improve the RA success probability of an IoT device by exploiting a distributed queue mechanism and then proposed an access resource grouping mechanism to reduce the access delay caused by the queuing process of the distributed queue mechanism. To increase RA success probability, the work in [4] proposed to increase the number of preambles at the first step of the RA procedure by utilizing a spatial group mechanism and improve resource utilization through non-orthogonally allocating uplink channel resources at the second step of the RA procedure. Additionally, without knowing the statistical characteristic of traffic, a reinforcement learning-based algorithm was proposed in [5] to determine the uplink resource configuration for RA such that the average number of served IoT devices was maximized while ensuring a high RA success probability.

Except for the IoT service, future networks are envisioned to simultaneously support different services and applications with significantly different requirements on reliability, latency and bandwidth. As a result, researchers are now paying more attention to the service multiplexing of IoT/mMTC and many other services such as eMBB and URLLC. For example, in [15] developed a two-level scheduling process to allocate dynamically dedicated bandwidth to each network slice according to workload demand and slices’ quality of service (QoS) requirement such that flexible resource allocation could be implemented. The work in [16] proposed to maintain slice-specific radio resource control elements with which the RAN protocol stacks and different slices were configured. Besides, the work in [17] aimed to optimize the virtual network functions and infrastructure resources such as the system bandwidth to implement slice recovery and reconfiguration for mMTC and eMBB service provision. The work in [18] proposed to maintain slice isolation between mMTC and eMBB slices and meet the performance requirements of these slices through limiting and dynamically updating the amount of resources allocated to each slice and monitoring the resource usage of each slice. After representing the slice performance requirements as the required amount of resources per deadline interval, an idea of earliest-deadline and first-scheduling was exploited in [19] to allocate radio resources to mMTC, eMBB, and URLLC slices effectively.

III. SYSTEM MODEL

We consider a coordinated-multipoint-enabled RAN slicing system for mIoT and bursty URLLC multiplexing service pro-
vision. From the viewpoint of the infrastructure composition, the system mainly includes one baseband unit (BBU) pool and multiple remote radio heads (RRHs) that connect to the BBU via fronthaul links. From the perspective of network slicing, two types of inter-slices, i.e., mIoT slices and URLLC slices, are exploited in this system with \( S^t \) and \( S^u \) representing the sets of mIoT slices and URLLC slices, respectively. We focus on the modelling of uplink IoT data transmission in mIoT slices and the modelling of downlink URLLC data transmission in URLLC slices. The IoT devices are spatially distributed in \( \mathbb{R}^2 \) according to an independent homogeneous Poisson point process (PPP) \( \Phi_R = \{ u_{i,s}; s \in S^t, i = 1, 2, \ldots \} \) with intensity \( \lambda^t_i \), where \( u_{i,s} \) denotes the location of the \( i \)-th IoT device in the \( s \)-th mIoT slice. There are also \( N^u \) URLLC devices that are randomly and evenly distributed in \( \mathbb{R}^2 \). The RRHs are spatially distributed in \( \mathbb{R}^2 \) according to an independent PPP \( \Phi_R = \{ v_{j}; j = 1, 2, \ldots \} \) with intensity \( \lambda_R \), where \( v_{j} \) represents the location of the \( j \)-th RRH. The number and locations of IoT devices and RRHs will be fixed once deployed. Besides, each RRH is equipped with \( K \) antennas, and each device is equipped with a single antenna. In IoT network slices, each IoT device is assumed to connect to its associated RRH. In URLLC network slices, RRHs cooperate to transmit signals to a URLLC device to improve its signal-to-noise ratio (SNR). A flexible frequency division multiple access (FDMA) technique is utilized to achieve the inter-slice and intra-slice interference isolation [21].

The system time is discretized and partitioned into time slots and minislots with a time slot consisting of \( T \) minislots. On the one hand, at the beginning of each time slot, a RAN slicing coordinator [22] will decide whether to accept or reject received network slice requests which will be defined in the following subsections. Once a slice request is accepted, a network slice management will be responsible for activating or creating a virtual slice that is well resource-configured to satisfy the QoS requirements of devices in the slice [23]. The slice configuration process is time costly and will generally be conducted in a timescale of minutes to hours [21]. On the other hand, at the beginning of each minislot, each active IoT device may try to connect to its associated RRH, and RRHs will generate cooperated beamformers based on sensed channel coefficients.

### A. mIoT slice model

By referring to the concept of a network slice [16], especially from the viewpoint of the QoS requirement of a slice, we can define a mIoT slice request as follows.

**Definition 1** (mIoT slice request). A mIoT slice request is defined as a tuple \( \{ \lambda^t_i, \gamma^t_i, N^t_{a,s} \} \) for any slice \( s \in S^t \), where \( \gamma^t_i \) is the requirement of data transfer rate from an IoT device in \( s \) to its associated RRH, \( N^t_{a,s} \) denotes the number of accumulated packets in a queue of an IoT device in \( s \).

In this paper, all mIoT slice requests are always accepted by the RAN slicing coordinator. IoT devices with the same data transfer rate are assigned to the similar slice. For an IoT device in \( s \), if it has the opportunity to send its endogenous arrival packets to the corresponding RRH, then it will randomly select a preamble (e.g., orthogonal Zadoff-Chu sequences) from a BBU-maintained preamble pool and transmit the preamble to the RRH at the data rate \( \gamma^t_i \). Just like the literature [23], [24], if the RRH can successfully decode the preamble, then a connection between the IoT device and the RRH is considered to be set up although the whole connection establishment process usually follows an RA four-step procedure [8]. In other word, the RA success probability is regarded as the probability of successfully transmitting a preamble in this paper. Next, we will analyze an IoT device queue evolution model, with the analysis of which the RA success probability of the IoT device will be derived.

**1) Queue evolution model:** The queue evolution process consists of the packet arrival process, packet accumulate process, and packet departure process.

During minislot \( t \), a Poisson distribution with intensity (or arrival rate) \( \epsilon_{w,s}(t) \) is exploited to model the random, mutually independent endogenous packet arrivals in an IoT device in slice \( s \). Then during minislot \( t \) with a duration \( \tau \), the arrival intensity of new packets can be expressed as \( \mu_{w,s}(t) = \epsilon_{w,s}(t) \tau \). Once arrived, new packets will not be sent out immediately in general and will enter a queue, which is modelled as an \( M/M/k \) queue with unlimited capacity, to wait for their scheduling. In the \( M/M/k \) queue, packets will be scheduled according to the first-come, first-served (FCFS) basis. The unlimited queue capacity indicates that the age of information [25], [26] of new arrivals will not be considered, and packets will not be dropped before sending out. Besides, owing to the RA behavior of a slotted-ALOHA protocol, new arrivals during \( t \) will only be counted at minislot \( t+1 \). Thus, the accumulated number of packets \( N^t_{a,s}(t) \) of a randomly selected IoT device in slice \( s \) at \( t \) is determined by the accumulated number of packets and the number of new arrivals at \( t-1 \) and whether the preamble of the device can be successfully decoded by its associated RRH. Table I shows the evolution of accumulated packets in an IoT device. In this table, \( x_{s} = \gamma^t_i/L \) packets at the head of the queue will be popped out if the corresponding RA succeeds, where \( L \) denotes the IoT packet length; otherwise, they will be kept in the queue and wait for the opportunity of re-transmission at the next minislot. The operation \( \lfloor x \rfloor^+ = \max(x, 0) \).

With the evolution of accumulated packets, we can define the non-empty probability of the queue of an IoT device in \( s \) as the following.
### Definition 2 (Non-empty probability). At minislot $t$, for a randomly selected IoT device in slice $s \in S^I$, the probability that its queue is not empty can be defined as

$$P_{ne,s}(t) = \mathbb{P}\{N_{s,a}(t) > 0\} \quad \forall s \in S^I$$

(1) implicitly reflects that new arrival packets at $t$ will not be sent out immediately. According to the evolution of $N_{s,a}(t)$, it can be observed that $P_{ne,s}(t)$ is determined by the probability distribution of $N_{s,a}(t-1)$ and the RA success probability. Since these probabilities and their correlations are unknown, the derivation of the explicit expression of $P_{ne,s}(t)$ is difficult.

Next, we describe the packet departure process combined with a frame and minislot structure for mIoT packets transmission. As mentioned above, partly because of the limitation on the frame and minislot structure, NB-IoT and LTE-M can only admit 50,000 devices. For NB-IoT, only one physical resource block (PRB) with a bandwidth of 180 KHz in the frequency domain is allocated for IoT transmission, and each physical channel occupies the whole PRB. For LTE-M, although the physical channels are time and frequency multiplexed, it only reserves six in-band PRBs with a total bandwidth of 1.08 MHz in the frequency domain for IoT data transmission. Therefore, the frame and minislot structure for mIoT transmission should be revisited if more RA requests from IoT devices want to be accepted.

Fig. 1 depicts a frame and minislot structure for mIoT transmission in each mIoT slice. In this structure, both the frequency division multiplexing scheme and code division multiplexing scheme are leveraged to admit more IoT devices in the way of alleviating the mutual device interference. Particularly, the frequency division multiplexing scheme alleviates signal interference through orthogonal frequency allocation, and the code division multiplexing scheme mitigates the co-channel signal interference via reducing the cross-correlation of simultaneous transmissions. The combination of the two schemes may significantly mitigate interference experienced at an RRH. In this way, the QoS requirements of more IoT devices may be satisfied, and the RAN slicing system may support more IoT devices. For a mIoT slice $s \in S^I$, each subframe includes $F_s$ orthogonal uplink physical RA channels (PRACHs). A single tone mode with a tone spacing of size $a$ MHz is adopted for each uplink PRACH, which indicates that each PRACH occupies a PRB. At the beginning of each minislot, an active IoT device, i.e., the device’s queue is non-empty, will randomly choose a preamble from a set of non-dedicated RA preambles of size $\xi$ and transmit the preamble through a randomly selected PRACH. For each preamble, it has an equal probability $\frac{1}{F_s}$ to be chosen by each IoT device. Similarly, each PRACH has an equal probability $\frac{1}{F_s}$ to be selected. Thus, the average number of IoT devices in mIoT slice $s \in S^I$ choosing the same PRACH and the same preamble is $\frac{\xi F_s}{2}$. Notably, a greater $F_s$ may significantly reduce signal interference experienced at each RRH.

Then, the following question should be tackled: how many PRBs should be reserved for mIoT transmission? To improve the resource utilization, the resource allocated to mIoT should be determined according to the requirements of mIoT and other coexistence services. It motivates us to optimize the resources orchestrated for the mIoT service except for analyzing the RACH procedure of IoT devices. The optimization procedure will be discussed in detail in the next section.

2) Access control scheme: In a mIoT network slice, as the slotted-ALOHA protocol allows all active IoT devices to request for RA at the beginning of each minislot without checking the status of channels, IoT devices may simultaneously transmit preambles. It may incur severe slice congestion that may lower the RA success probabilities of IoT devices and degrade the system performance. Access control has been considered as an efficient proposal of alleviating congestion, and many access control schemes have been proposed [7]. In this paper, we aim at illustrating the performance difference between a network slicing system without access control and with access control. Therefore, we adopt the following two schemes [7]:

- **Unrestricted scheme:** each active IoT device requests the RACH at the beginning of minislot $t$ without access restriction. If mIoT slices are not crowded or in a lightly crowded condition, then this scheme may result in high packet queuing delay.

- **Access class barring (ACB) scheme:** at the beginning of $t$, each active IoT device throws a random number $q \in [0, 1]$ and can request the RACH only if $q < P_{ACB}$, where $P_{ACB}$ is an ACB factor determined by RRHs based on the slice congestion condition. The ACB scheme can relieve slice congestion to some extent by reducing RACH requests of active IoT devices.

With the introduced access control schemes, we can define the non-restriction probability of a randomly selected IoT device in $s$ as follows.

### Definition 3 (Non-restriction probability). At minislot $t$, for a randomly selected IoT device in slice $s \in S^I$, the probability that its RACH request is not restricted is defined as

$$P_{nr,s}(t) = \mathbb{P}\{\text{Unrestricted RACH requests}\} \quad \forall s \in S^I$$

(2) For all $s \in S^I$ at any minislot $t$, we have $P_{nr,s}(t) = 1$ for the unrestricted scheme and $P_{nr,s}(t) = P_{ACB}$ for the ACB.
scheme.

3) Analysis of RA success probability: For an RRH, two significant reasons may lead to an error preamble decoding i) the achieved preamble transfer rate at the RRH is less than a preset threshold; ii) the RRH simultaneously decodes at least two similar co-channel preambles, and thus preamble collision occurs. The research of the mitigation of preamble collision has been well conducted in [11], [28]. Just like [29], we focus on the exploration of enabling successful single preamble transmission that is discussed in detail as follows.

We utilize a power-law path-loss model to calculate the path-loss between an IoT device and its RRH in mIoT slices and utilize a truncated channel inversion power control scheme to eliminate the ‘near-far’ effect. In the power-law path-loss model, the IoT device transmit power decays at the rate of $r^{-\varphi}$ with $r$ representing the propagation distance and $\varphi$ denoting the path-loss exponent. In the power control scheme, IoT devices associated with the same RRH compensate for the path-loss to maintain the average received signal power at the RRH equal to a threshold $\rho_o$. Without loss of generality, the cutoff threshold $\rho_o$ is set to be the same for all RRHs. Owing to the channel deep fading, severe co-channel interference, and insufficient transmit power, an IoT device may experience uplink preamble transmission outage. The following definition describes the definition of the probability that a randomly selected IoT device can successfully transmit a chosen preamble to its corresponding RRH.

**Definition 4 (RA success probability).** At minislot $t$, for a randomly selected active IoT device in slice $s \in S^t$, its RA success probability is defined as

$$P_s(t) = \mathbb{P}\{r_s(t) \geq r_{s|}^{th}\}, \forall s \in S^t$$

where $r_s(t) = a \log_2(1 + S I N R_s(t))$ denotes the achieved preamble transfer rate at the IoT device’s associated RRH and $S I N R_s(t)$ is the signal-to-interference-plus-noise ratio (SINR).

Then, for any active IoT device in $s$, its QoS requirement is given by

$$P_s(t) \geq \pi_s, \forall s \in S^t$$

where $\pi_s$ denotes a threshold of the required RA success probability.

This definition shows that the QoS requirement of each active IoT device in $s$ should be satisfied if the slice request of $s$ is accepted. The definition also states that $P_s(t)$ is correlated with the non-empty probability $P_{ne,s}(t)$. Recall that the RA success probability of an IoT device impacts its non-empty probability, we can know that the RA success probability and the non-empty probability are intertwined. Additionally, $S I N R_s(t)$ is a function of complicated co-channel interference. Thus, it is hard to obtain the closed-form expression of $P_s(t)$.

Without any loss in generality, we perform the analysis of RA success probability on an RRH located at the origin. According to Slivynak’s theorem [30], the analysis holds for a generic RRH located at a generic location. For a randomly selected IoT device with non-empty queue in $s \in S^t$, the theoretical preamble transfer rate experienced at the RRH located at the origin can take the form

$$r_s(t) = a \log_2 \left(1 + \frac{\rho_o h_o}{\sigma^2 + I_s(t)}\right), \forall s \in S^t$$

where $\sigma^2$ represents the noise power, $I_s(t)$ denotes signal interference received at the RRH, the useful signal power equals to $\rho_o h_o$ due to the truncated channel inversion power control\(^2\) with $h_o$ denoting the channel power gain between the IoT device and the RRH. It is noteworthy that the channel power gain experienced at a generic RRH is related to the spatial locations of both the RRH and its associated IoT devices. Nevertheless, we drop the spatial indices for notation lightening. Besides, just like [31], all channel gains are assumed to be known and be independent of each other, independent of the spatial locations, symmetric and are identically distributed (i.i.d.). Considering both the particular IoT device deployment environment and the convenience of theoretical analysis, the Rayleigh fading is assumed, and the channel power gain $h_o$ is assumed to be exponentially distributed with unit mean.

Based on the following five facts, we next present the analytical expression of signal interference

- **Fact 1**: the average signal received from any single IoT device belonging to inter-cells is strictly less than $\rho_o$.
- **Fact 2**: the average interference signal received from any single interfering IoT device associated with the origin RRH strictly equals to $\rho_o$.
- **Fact 3**: IoT devices choosing the same co-channel preamble as the randomly selected IoT device may become an interfering IoT device.
- **Fact 4**: at each minislot, IoT devices with non-empty queue may become interfering IoT devices.
- **Fact 5**: IoT devices in difference slices may not mutually interfere.

Note that Fact 1 and Fact 2 are direct consequences of the device-RRH association policy and power control scheme. Fact 5 holds due to the exploration of intra-slice isolation. Therefore, the aggregate interference received at the origin RRH can take the following form

$$I_s(t) = \sum_{m \in \Phi_s \setminus \{o\}} \mathbb{I}(p_m ||d_m||^{-\varphi} = \rho_o) \mathbb{I}(N_o,s(t) > 0) \times \mathbb{I}(f_m = f_o) \rho_o h_m, \forall s \in S^t$$

where $o$ represents the randomly selected IoT device associated with the RRH at the origin, $p_m$ represents the transmit power of the $m$-th IoT device, $||d_m||$ is the distance between the $m$-th IoT device and the origin RRH, $f_o$ denotes the preamble and channel chosen by the randomly selected IoT device. $f_o = f_m$ indicates that the randomly selected IoT device and the $m$-th IoT device select the same preamble and channel. $\mathbb{I}(\cdot)$ is the indicator function that equals to one if the statement $\mathbb{I}(\cdot)$ is true; otherwise, it equals to zero. Just like [32], in (6), co-channel inter-cell interference is assumed.

\(^2\)Owing to the truncated channel inversion power control, not all of the IoT devices in mIoT slices can communicate in the uplink when the cutoff threshold is relatively high [31]. However, this paper assumes that the transmit power of each IoT device is large enough such that the IoT device will not experience preamble outage resulting from the insufficient power.
as a part of thermal noise mainly because of the severe wall penetration loss.

Then, for the randomly selected IoT device in $s \in S^I$, we can rewrite (5) as the following form with (6).

$$
P_s(t) = \mathbb{P}\{SINR_{a,s}(t) \geq \theta^h_s\}
= \mathbb{P}\{h_o \geq \theta^h_s \rho_o / (\sigma^2 + I_s(t))\}
= \exp\left(-\frac{\theta^h_s}{\rho_o / (\sigma^2 + I_s(t))}\right), \forall s \in S^I
$$

where $\theta^h_s = 2\gamma^h_s / \rho_o - 1$. (a) follows from the full probability law over $I_s(t)$, and $I_{\mathcal{L}_s}(\cdot)$ denotes the Laplace transform (LT) of the probability density function (PDF) of the random variable $I_s(t)$. Note that the notation $I_{\mathcal{L}_s}(\cdot)$ is a terminology that is a slight abuse of subscript $I_s(t)$.

The following lemma characterizes the LT of aggregate interference $I_s(t)$.

**Lemma 1.** For the origin RRH, the LT of its received aggregate interference from active IoT devices associated with it is given by

$$
I_{\mathcal{L}_s}(\psi_s) = \frac{1 + \psi_s \rho_o}{(1 + \alpha_s \psi_s \rho_o / (1 + \psi_s \rho_o))^3.5} - \frac{1 + \psi_s \rho_o}{(1 + \alpha_s)^3.5}
$$

where $\psi_s = \theta^h_s / \rho_o$, $\alpha_s = \frac{P_{ne,s}(t)P_{ne,s}(t)\lambda_1s_{ta,s}}{5.5\lambda s_{ta,s}}$, for all $s \in S^I$.

**Proof.** Please refer to Appendix A.

With the conclusion in Lemma 1 we can then obtain the mathematical expression of the RA success probability of a randomly selected IoT device at $t$ in the following corollary.

**Corollary 1.** For a randomly selected IoT device in a mIoT slice $s \in S^I$, its RA success probability at minislot $t$ is given by

$$
P_s(t) = \frac{1 + \psi_s \rho_o e^{-\psi_s \sigma^2}}{(1 + \alpha_s \psi_s \rho_o / (1 + \psi_s \rho_o))^3.5} - \frac{1 + \psi_s \rho_o e^{-\psi_s \sigma^2}}{(1 + \alpha_s)^3.5}
$$

**Proof.** By substituting (8) into (7), we can obtain (9).

Although Corollary 1 presents a mathematical expression of $P_s(t)$, the expression is not in the closed-form as it is a function of $P_{ne,s}(t)$ the closed-form expression of which is not obtained. Next, we derive the closed-form expression of $P_{ne,s}(t)$.

4) Analysis of non-empty probability: According to the definition of non-empty probability, $P_{ne,s}(t)$ is correlated with the number of accumulated packets $N_{a,s}(t)$ of the randomly selected IoT device in mIoT slice $s$ thus, we theoretically analyze the non-empty probability of the randomly selected IoT device as the following.

As the number of the accumulated packets in the queue of a randomly selected IoT device in slice $s \in S^I$ at the 1st minislot is empty, its non-empty probability $P^1_{ne,s}$ at the 1st minislot can take the form

$$
P^1_{ne,s} = \mathbb{P}\{N^1_{a,s} > 0\} = 0, \forall s \in S^I
$$

where we write $x^I$ instead of $x(t)$ to lighten the notation.

The following lemma presents the closed-form expression of the non-empty probability of a randomly selected IoT device served by the origin RRH when minislot $t > 1$.

**Lemma 2.** The number of accumulated packets of a randomly selected IoT device served by the origin RRH at minislot $t > 1$ may be approximately Poisson distributed. Therefore, for any mIoT slice $s \in S^I$, we approximate the number of accumulated packets $N^t_{a,s}$ at minislot $t$ as a Poisson distribution with intensity $\mu^t_{a,s}$, which is given by

$$
\mu^t_{a,s} = \left[\mu^t_{w,s} + \mu^t_{a,s} - P_s(t - 1) \left(1 - e^{-\mu^t_{w,s} - \mu^t_{a,s}}\right)\right]^{+}, \forall s \in S^I
$$

**Proof.** Please refer to Appendix B.

Combine with (9) and (11), the closed-form expression of $P_s(t) (s \in S^I)$ can be obtained.

**B. Bursty URLLC slice model**

Similar to the definition of a mIoT slice request, a bursty URLLC slice request can be defined as the following.

**Definition 5** (Bursty URLLC slice request). A bursty URLLC slice request is defined as four tuples $\{I^o_s, D_s, \alpha, \beta\}$ for slice $s \in S^a$, where $I^o_s$ denotes the number of URLLC devices in $s$, $D_s$ denotes the transmission latency requirement of each URLLC device in $s$, $\alpha$ and $\beta$ represent the packet blocking probability and the codeword error decoding probability of each URLLC device, respectively.

In this definition, URLLC devices are grouped into $|S^a|$ clusters according to the transmission latency requirement of each device. Owing to the low latency requirement URLLC packets should be immediately scheduled upon arrival; thus, all URLLC slice requests will always be accepted by the RAN coordinator. Except for the low packet error decoding probability that has been emphasized for URLLC transmission in a plenty of works [21], this paper attempts to orchestrate slice resources to reduce the packet blocking probability for bursty URLLC transmission. This is because the bursty characteristic of URLLC traffic [22] may lead to the packet blocking in URLLC slices, which may significantly reduce the reliability of URLLC transmission. Therefore, the indicators $\alpha$ and $\beta$ are involved to reflect the ultra-reliable requirement of URLLC transmission jointly.

Then, we address the following question: how to orchestrate slice resources for reducing packet blocking probability and codeword error decoding probability?

1) Reduction of packet blocking probability: As mentioned above, the bursty feature of URLLC traffic is the crucial factor that leads to the URLLC packet blocking for URLLC transmission. Therefore, we next model the bursty URLLC traffic based on which we discuss how to orchestrate slice resources to alleviate the impact of bursty URLLC traffic.
During minislots $t$, an independent homogeneous Poisson distribution with intensity $\lambda = \{\lambda_s; s \in S_u\}$ is utilized to model the number of bursty URLLC packets aggregated at RRHs, where $\lambda_s$ denotes the intensity of new arrivals destined to devices belonging to URLLC slice $s$.

Once arrived, new URLLC arrivals will enter a queue maintained by an RRH to be immediately scheduled. An M/M/W queueing system with limited bandwidth $W_u$ is exploited to model the queue. Without loss of any generality, we assume that each RRH maintains the same queue due to the exploration of cooperated transmission. In the queue, a packet destined to URLLC device $i \in I_s^u$, $s \in S_u$ will be allocated with a block of system bandwidth $\omega_{i,s}(t)$ for a period of time $d_s \leq D_s$ at $t$. Owing to stochastic variations in the bursty packet arrival process, the limited bandwidth may not be enough to serve new arrivals occasionally. As such, URLLC packet blocking may happen. To reduce the probability of URLLC packet blocking, the redesign of URLLC frame and minislots structure may be required.

At minislots $t$, let $P_b(\omega^u(t), \lambda, d, W^u(t))$ denote the packet blocking probability experienced at an RRH, where $\omega^u(t) = \{\omega_{i,1}^u(t), \ldots, \omega_{i,n}^u(t), \ldots, \omega_{i,n_i}^u(t)\} | S_u^i | S_u^i(t)$ and $d = \{d_1, \ldots, d_s, \ldots, d_{i(S_u^i)}\}$. The Theorem 1 in [33] provides us with a clue of redesigning the URLLC frame and minislots structure in the time-frequency plane for bursty URLLC traffic transmission. This theorem indicates that if we narrow the PRB of the URLLC frame in the frequency domain while widening it in the time domain, then the number of concurrent transmissions will be increased. As a result, the packet blocking probability is reduced.

Therefore, for a URLLC packet destined to device $i \in I_s^u$, $s \in S_u$, we should scale up $d_s$ and choose $d_s$ and $\omega_{i,s}^u(t)$ at $t$ using the following equation

$$d_s = D_s \text{ and } \omega_{i,s}^u(t) = \frac{b_i^u(t) r_{i,s}^u(t)}{\kappa D_s}, \forall i \in I_s^u, s \in S_u.$$ (13)

where $r_{i,s}^u(t)$ denotes channel uses for transmitting a URLLC packet, $\kappa$ is a constant reflecting the number of channel uses per unit time per unit bandwidth of FDMA frame structure and numerology, $b_i^u(t)$ is an indicator variable reflecting whether the QoS requirement of device $i$ in slice $s$ can be satisfied at $t$. As network resources are limited and shared by all network slices, not all URLLC devices can be guaranteed to be served at every minislots. Certainly, we can adjust the slice priority weight that will be introduced in the following section to guide the resource orchestration for enforcing the entire URLLC devices coverage.

Based on the result in [13] and the conclusion of the Lemma 3.2 in our previous work [22], we can derive the minimum upper bound of bandwidth orchestrated for URLLC slices in the following lemma.

**Lemma 3.** At minislots $t$, for a given M/M/W queue with packet arrival intensity $\lambda$ and a family of packet transmit rates $\{\kappa(t) r_{i,s}^u(t)\}$, let $W^u(r(t))$ denote the minimum upper bound of bandwidth orchestrated for URLLC slices such that $P^M_{Q/W^u} \leq \varsigma$ and $P_b(\omega^u(t), \lambda, D, W^u(r(t)))$ is of the order of $\alpha$, where $P^M_{Q/W^u}$ represents the queuing probability, and $D = \{D_1, \ldots, D_{|S_u|}\}$. If $\varsigma > \alpha$, then we have

$$W^u(r(t)) \approx \sum_{s \in S_u} \sum_{t \in I_u^s} \lambda_s b_i^u(t) r_{i,s}^u(t) +$$

$$\frac{\alpha - \varsigma}{\varsigma - \alpha} \min_{s \in S_u} \{\lambda_s D_s\} \left( \sum_{s \in S_u} \sum_{t \in I_u^s} \lambda_s b_i^u(t) r_{i,s}^u(t)^2 \right) \left( \sum_{s \in S_u} \sum_{t \in I_u^s} \lambda_s b_i^u(t) r_{i,s}^u(t)^2 \right)^{\frac{1}{2}}$$

(14)

**Proof.** We omit the proof here as the similar proof can be found in the proof section of Lemma 3.2 in [22].

2) Reduction of codeword error decoding probability:

The crucial factor that impacts the codeword error decoding probability is the network capacity. Next, we discuss the relationship between the network capacity and codeword error decoding probability.

For any URLLC slice $s \in S_u^u$, during minislot $t$, let $x_{i,s}^u(t)$ be the original data symbol destined to a URLLC device $i \in I_s^u$ with $E[|x_{i,s}^u(t)|^2] = 1$, $g_{i,s}(t) \in C^K$ be the transmit beampformer pointing at the device $i$ from the $j$-th RRH and $h_{i,j,s}(t) \in C^K$ be the channel coefficient between the $i$-th URLLC device and the $j$-th RRH. The channel coefficient may change over minislots. However, it is assumed to be i.i.d. over each minislot and remain unchanged during each minislot. Then, the received signal at device $i$ in $s$ during minislot $t$ is given by

$$\hat{x}_{i,s}^u(t) = \sum_{j \in J} h_{i,j,s}(t) g_{i,s}(t) x_{i,s}^u(t) + \sigma_{i,s}(t), \forall i \in I_s^u, s \in S_u$$

(15)

where the first term is the useful signal for $i$ and $\sigma_{i,s}(t) \sim \mathcal{CN}(0, \sigma^2_{i,s})$ is the additive white Gaussian noise (AWGN) experienced at $i$. Similar to [21], interference signal is not involved in (15) due to the utilization of a flexible FDMA mechanism. Then the SNR received at device $i$ in $s$ at minislot $t$ can be written as

$$SNR_{i,s}^u(t) = \frac{|\sum_{j \in J} h_{i,j,s}(t) g_{i,s}(t)|^2}{\phi \sigma^2_{i,s}}, \forall i \in I_s, s \in S_u$$

(16)

where $\phi > 1$ is an SNR loss coefficient. The perception of perfect channel status information (CSI) or accurate channel coefficients requires the information exchange between an RRH and its associated device before data transmission, the process of which is generally time consuming. URLLC packets, however, have a stringent latency requirement. As a result, perfect CSI or accurate channel parameters may be unavailable for URLLC transmission, which may incur the SNR loss. The coefficient $\phi$ is then utilized to characterize the SNR loss [34].

Shannon capacity formula is created under a crucial assumption of transmitting a block with long enough blocklength. However, URLLC packets are typically very short to satisfy the ultra-low latency requirement. Thus, the famous Shannon capacity formula cannot be utilized to model the URLLC transmission data rate and capture the corresponding codeword error decoding probability. For URLLC transmission, we resort to the capacity analysis for a finite blocklength channel coding regime derived in [35]. For any device $i \in I_s^u$, $s \in S_u$, the
number of transmitted information bits $L_{t,i,s}(t)$ at minislot $t$ using $r_{i,s}(t)$ channel uses in AWGN channel can be accurately correlated with the codeword error decoding probability $\beta$ according to the following equation

$$\begin{align*}
L_{t,i,s}(t) & \approx r_{i,s}(t)C(SNR_{t,i,s}(t)) - Q^{-1}(\beta) \sqrt{r_{i,s}(t)V(SNR_{t,i,s}(t))}, \forall i \in I_s, s \in S^n
\end{align*}$$

where $C(SNR_{t,i,s}(t)) = \log_2(1 + SNR_{t,i,s}(t))$ is the AWGN channel capacity under infinite blocklength assumption, $V(SNR_{t,i,s}(t)) = \ln^2 2 \left(1 - \frac{1}{1 + SNR_{t,i,s}(t)}\right)$ is the channel dispersion, $Q(\cdot)$ is the $Q$-function. It is noteworthy that a URLLC packet will usually be coded before transmission and the generated codeword will be transmitted in the air interface such that the transmission reliability can be improved.

The complicated expression of $V(SNR_{t,i,s}(t))$ in (17) significantly hinders the theoretical analysis of network resources orchestrated for URLLC slices. Fortunately, as $V(SNR_{t,i,s}(t))$ is maximized by $\ln^2 2$, the closed-form expression of the minimum upper bound of $r_{i,s}(t)$ ($i \in I_s, s \in S^n$) with a codeword error decoding probability $\beta$ can be given by

$$\begin{align*}
r_{i,s}(t) & = \frac{L_{t,i,s}(t) - \frac{Q^{-1}(\beta)^2}{2C(SNR_{t,i,s}(t))^2}}{\frac{Q^{-1}(\beta)^2}{2C(SNR_{t,i,s}(t))^2} \sqrt{1 + \frac{4L_{t,i,s}(t)C(SNR_{t,i,s}(t))}{(Q^{-1}(\beta))^2}}} \quad (18)
\end{align*}$$

IV. PROBLEM FORMULATION

This section aims to formulate the problem of RAN slicing for mIoT and bursty URLLC service multiplexing based on the above models.

In mIoT slices, each RRH may transmit feedback signal to its connected IoT devices for the connection establishment according to an RA four-step procedure [8]. Meanwhile, in URLLC slices, each RRH may transmit URLLC packets to URLLC devices. As the transmit power $E_j$ of each RRH is limited, we have the following transmit power constraint

$$\sum_{s \in S} \left(1 + \alpha_g \lambda_s^{1} E_j\right) + \sum_{s \in S^n \setminus I_s} \sum_{i \in I_s} b_{i,s}(t)g_{ij,s}(t)g_{ij,s}(t) \leq E_j, \forall j \in J \quad (19)$$

where $E_j$ is assumed to be a constant and denotes the transmit power of the $j$-th RRH for connecting to its associated IoT devices over downlink. $\alpha_g$ represents a safety margin coefficient. As a PPP with intensity $\lambda_s^{1}$ is utilized to model the distribution of IoT devices, the actual number of IoT devices may be greater than $\lambda_s^{1}$ once deployed. As a result, the coefficient $\alpha_g$ is introduced to reserve transmit power for exceed IoT devices.

In the RAN slicing system, as the total limited system bandwidth $W$ will be shared by mIoT slices and URLLC slices, we have the following bandwidth constraint

$$\sum_{s \in S^I} \left(1 + \alpha_g \omega_s(\bar{t})\right) + W^u(r(\bar{t})) \leq W \quad (20)$$

where $\omega_s(\bar{t})$ denotes the bandwidth allocated to mIoT slice $s \in S^I$ that is correlated with $F_s$ by means of $F_s = \frac{\omega_s(\bar{t})}{\alpha_j}$, and $\alpha_g \omega_s(\bar{t})$ denotes a block of reserved bandwidth.

In (20), $F_s$ is an integer, and some integer variable recovery schemes [36] can be leveraged to obtain the suboptimal $F_s$. However, considering the high computational complexity of optimizing an integer variable and the utilization of the scheme of spectrum safety margin, we directly relax the integer variable into a continuous one, i.e., let $F_s = \omega_s(\bar{t})/\alpha_j$. Without loss of any generality, we regard $\omega_s(\bar{t})$ as an independent variable below. Besides, as at least one PRB should be allocated to mIoT slices, we have

$$\omega_s(\bar{t}) \geq a, \forall s \in S^I \quad (21)$$

Owing to the exploration of mIoT and bursty URLLC service multiplexing, we should orchestrate network resources for all mIoT slices and URLLC slices to simultaneously maximize the utilities of mIoT slices and URLLC slices.

For a mIoT slice $s \in S^I$, its primary goal is to offload as many data packets as possible from IoT devices. In this way, the number of accumulated packets in each IoT device should be kept at a low level. Considering that a great RA success probability of an IoT device will lead to a low number of accumulated packets in its queue, we define the utility of a mIoT slice as the following.

**Definition 6** (mIoT slice utility). Over a time slot of duration $T$, the mIoT slice utility is defined as the time-average of RA success probabilities of IoT devices in all mIoT slices, which is given by

$$\bar{U}^I = \frac{1}{T} \sum_{t=1}^{T} U^I(t) = \frac{1}{T} \sum_{t=1}^{T} \tilde{P}(t) \quad (22)$$

where $\tilde{P}(t) = \frac{\lambda_l^I P_s(t)}{\sum_{s \in S^I} \lambda_l^I}$ with the numerator $\lambda_l^I P_s(t)$ representing the expected sum of RA success probabilities of IoT devices in slice $s \in S^I$ and the denominator $\sum_{s \in S^I} \lambda_l^I$ denoting a normalization coefficient.

In (22), $\frac{\lambda_l^I}{\sum_{s \in S^I} \lambda_l^I}$ can be regarded as an intra-slice priority coefficient. A mIoT slice serving more IoT devices will be orchestrated with more network resources.

For a URLLC slice $s \in S^n$, its primary objective is to maximize the slice gain that is reflected by the parameters in the slice request at a low cost. Therefore, we define an energy-efficient utility for URLLC slices, as presented below.

**Definition 7** (Bursty URLLC slice utility). Over a time slot of duration $T$, the bursty URLLC slice utility is defined as the time-average energy efficiency for serving all URLLC devices, which is given by

$$\bar{U}^u = \frac{1}{T} \sum_{t=1}^{T} U^u(t) = \frac{1}{T} \sum_{t=1}^{T} \sum_{s \in S^n} U^u(D_s(t)g_{ij,s}(t))$$

$$= \frac{1}{T} \sum_{t=1}^{T} \sum_{s \in S^n} \sum_{i \in I_s} \sum_{j \in J} b_{i,s}(t)g_{ij,s}(t)g_{ij,s}(t) \quad (23)$$

where $\eta$ is an energy efficiency coefficient reflecting a tradeoff between the URLLC slice gain and the RRH energy consumption.
Then, over a time slot of duration $T$, the RAN slicing problem for mIoT and URLLC service multiplexing can be formulated as follows

$$\max b^*_i(t), \omega_s(t), \psi_{i,s}(t) \quad \text{subject to:}$$

$$b^*_i(t) \in \{0, 1\}, \forall s \in S^u, i \in I^u$$

$$\text{constraints (11), (19) – (21) are satisfied.}$$

where $\hat{\rho}$ is an inter-slice priority coefficient reflecting the priority of orchestrating network resources for mIoT slices and URLLC slices.

The mitigation of (24) is quite challenging mainly because

- **indeterministic objective function**: (24) should be optimized at the beginning of the 1st minislot. The time-averaged objective function of (24) can only be exactly computed according to the future channel information. Therefore, the value of the objective function is indeterministic at the beginning of the 1st minislot.

- **two timescale issue**: the creation of a network slice is performed at a timescale of time slot. Thus, the variable $\omega_s(t)$ should be determined at the beginning of the time slot $t$ and kept unchanged over the whole time slot. The channel, however, is time-varying. As a result, the beamformer $g_{i,j,s}(t)$ should be optimized at each minislot. In summary, the variables in (24) should be optimized at two different timescales.

- **thorny optimization problem**: at each minislot $t$, the constraint (4) is non-convex over $\omega_s(t)$, and the constraints (19), (20) are non-convex over $g_{i,j,s}(t)$, which together lead to a non-convex problem.

V. PROBLEM SOLUTION WITH SYSTEM GENERATED CHANNEL

This section aims to tackle these challenges by exploiting of an SAA technique [37], an ADMM method [38], a semidefinite relaxation scheme and a Taylor expansion scheme.

A. Sample average approximation and alternating direction method of multipliers

As mIoT slices and URLLC slices share the network resources, both $U^i$ and $U^u$ may be determined by channel coefficients experienced by URLLC slices. At each minislot $t$, due to the i.i.d. assumption on the channel coefficients of URLLC slices, we have

$$\frac{1}{T} \sum_{t=1}^{T} U^i(t) + \frac{1}{T} \sum_{t=1}^{T} \hat{\rho} U^u(t) \approx E_h \left[ \tilde{U}^i + \hat{\rho} \tilde{U}^u \right]$$

where $h$ represents the channel samples of URLLC slices collected at the beginning of the time slot $t$.

Given a collection of channel samples $\{h_m\}$ with $h_m = [h_{11, m; \ldots; h_{1J, m; \ldots; h_{N+1, m}, S^u = m}]$ and $m \in M = \{1, \ldots, M\}$, Just like (22), as constraints (24b) and (24c) construct a nonempty compact set, the conclusion of Proposition 5.1 in (22) is applicable to this paper by exploiting the SAA technique. The conclusion indicates that if the number of channel samples $M$ is reasonably large, then $\frac{1}{M} \sum_{m=1}^{M} U^i_m + \hat{\rho} \frac{M}{T} \sum_{m=1}^{M} U^u_m$ converges to $E_h \left[ \tilde{U}^i + \hat{\rho} \tilde{U}^u \right]$ uniformly on the nonempty compact set almost surely. In other words, the SAA technique enables us to use the channel samples collected at the beginning of a time slot to approximate the unknown channel coefficients over the time slot. For notation lightening, we write $x_m$ instead of $x(m)$ that represents a variable corresponding to the channel sample $h_m$.

Recall that the variable $\omega_s(t)$ will be kept unchanged over the time slot $t$ and the beamformer $g_{i,j,s}(t)$ should be calculated at each minislot $t$, we can further consider (24) as a global consensus problem, which can be effectively mitigated by an ADMM method. In this problem, $\omega_s(t)$ is a global consensus variable that should be maintained in consensus for all $h_m$ and $g_{i,j,s}$ that is calculated based on $h_m$ is a local variable.

The fundamental principle of an ADMM method is to impose augmented penalty terms characterizing global consensus constraints on the objective function of an optimization problem. In this way, the local variables can be driven into the global consensus while still attempting to maximize the objective function. Let $G_{i,s,m} = g_{i,s,m}I_{i,s,m} \in \mathbb{R}^{J \times K \times J \times K}$, $H_{i,s,m} = h_{i,s,m}I_{i,s,m} \in \mathbb{R}^{J \times K \times J \times K}$, where $g_{i,s,m} = [g_{i1, s, m; \ldots; g_{iJ, s, m}] \in \mathbb{C}^{JK \times 1}$ and $h_{i,s,m} = [h_{i1, s, m; \ldots; h_{iJ, s, m}] \in \mathbb{C}^{JK \times 1}$. By applying the matrix property $G_{i,s,m} = g_{i,s,m}I_{i,s,m} \Leftrightarrow G_{i,s,m} \succeq 0, \text{rank}(G_{i,s,m}) \leq 1$ and utilizing the conclusions of SAA and ADMM, we can approximate (24) as the following problem at the beginning of the time slot $t$

$$\min_{\{\omega_s, \omega_s(t), b^*_i, \psi_{i,s}, G_{i,s,m}\}} \sum_{m=1}^{M} \left[ \frac{U^i_m}{M} - \frac{\hat{\rho} U^u_m}{M} \right] +$$

$$\sum_{m=1}^{M} \sum_{s \in S^r} \left[ \psi_{s} (\omega_s - \omega_s(t)) + \frac{\mu}{2} \|\omega_s - \omega_s(t)\|^2 \right]$$

subject to:

$$P_{s,m} \geq \pi_s, \forall s \in S^r, m \in M$$

$$\sum_{s \in S^r} (1 + \alpha_g) \frac{\lambda_s}{\lambda_R} E_j + \sum_{s \in S^u, i \in I^u} b^u_{i,s,m} (Z_j G_{i,s,m}) \leq E_j$$

$$\forall j \in J, m \in M$$

$$\sum_{s \in S^r} (1 + \alpha_g) \omega_s(t) + W^u (r_m) \leq W, m \in M$$

$$\forall s \in S^u, i \in I^u, m \in M$$

$$\text{rank}(G_{i,s,m}) \leq 1, \forall s \in S^u, i \in I^u, m \in M$$

$$b^u_{i,s,m} \in \{0, 1\}, \forall s \in S^u, i \in I^u, m \in M$$

where $\psi_{s,m}$ is the Lagrangian multiplier, $\mu$ is a penalty coefficient, $Z_j$ is a square matrix with $J \times J$ blocks, and each block in $Z_j$ is a $K \times K$ matrix. In $Z_j$, 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.

(24) is now reduced to a deterministic single timescale problem (26). What is more, (26) can be split into $M$.
separate problems that can be optimized in parallel as its computational complexity that may be impractical in implementation. Therefore, a greedy scheme of the computation complexity (27) is quite difficult. We next discuss how to handle this hard problem.

2) joint bandwidth and beamforming optimization: Given the obtained \( \bar{b}_{i,sm} \), (26) will be reduced to the following joint bandwidth and beamforming problem.

\[
\{ \omega^{(k+1)}_{sm}, G^{(k+1)}_{i,sm} \} = \underset{\{ \omega_{sm}, \bar{b}_{i,sm}, \} }{\text{argmin}} \mathcal{L}(\omega_{sm}, G_{i,sm}) \tag{32a}
\]

subject to:

for the \( m \)-th sample, (26c) – (26f) are satisfied. (32b)

In (32), the low-rank constraint (26) is non-convex, and its objective function is not convex and even not quasi-convex w.r.t. \( \omega_{sm} \), the tackling of which is quite tricky.

To tackle the non-convex low-rank constraint (26), we resort to the semidefinite relaxation technique. The primary procedures of SDR are i) directly drop the low-rank constraint; ii) solve the optimization problem without the low-rank constraint to obtain the solution; iii) owing to the relaxation, the obtained solution cannot satisfy the low-rank constraint in general. If it is, its principal component is the optimal solution to the problem. If not, then some manipulations such as randomization/scale [39] are needed to perform on the solution to impose the low-rank constraint.

For the tricky objective function, we are reminded of the art of dealing with a non-convex function, i.e., study the structure of the function if it is non-convex. A crucial observation is that \( P_{sm} \) is quasi-concave w.r.t. \( \omega_{sm} \) although the objective function is not quasi-convex w.r.t. \( \omega_{sm} \). Therefore, we resort to the Taylor expansion to approximate the tricky objective function.

The following analysis is built under the following two facts

- the value of the objective function of (32) is mainly determined by that of \( P_{sm}^m(k) \) (or \( U_{sm}^m(k) \));
- for all \( s \in S^u \), \( m \in M \), the solution \( \omega_{sm} \) maximizing \( P_{sm}^m \) must locate in the range of \( [\omega_{sm}^l, S^*_{sm}] \) shown in Fig. 2 where \( \omega_{sm}^l \) denotes the lower bound of \( \omega_{sm} \) satisfying the constraint (26b), \( S^*_{sm} \) is the \( \omega_{sm} \) maximizing \( P_{sm} \), and the notation \( P_{sm}|\omega_{sm} \) is utilized to explicitly indicate that \( P_{sm} \) is a function of \( \omega_{sm} \).

Fact 1 holds because the linear terms w.r.t. \( \omega_{sm} \) will dominate the objective function as the consensus constraint is active. Besides, the quadratic terms pull local values towards the consensus; thus, they will also donate little to the objective function. Fact 2 holds because the total bandwidth is limited and shared. For example, given a value \( \omega_{sm,1} \in [S^*_{sm} + \delta_w, W] \) with \( \delta_w \) being a small positive constant, there must exist a value \( \omega_{sm,2} \in [\omega_{sm}^l, S^*_{sm}] \) such that \( P_{sm}|\omega_{sm,1} = P_{sm}|\omega_{sm,2} \).

Thus, a small \( \omega_{sm} \) will be preferred as it indicates that more bandwidth can be allocated to URLLC slices to further improve the objective function.

For all \( s \in S^u \), it can be proved that \( P_{sm} \) is concave in the interval \( (a_1, a_2) \) by evaluating the second-order derivative
of $P_{sm}$. Therefore, we can leverage the 2nd degree Taylor expansion to approximate $P_{sm}$ in this interval. Considering that $P_{sm}$ is convex in the interval $[\omega_{sm}^b, a_1]$, the 1st degree Taylor expansion is always leveraged to obtain the lower bound of $P_{sm}$. However, this interval is usually rather narrow, and the value of $P_{sm}$ in this interval is much lower than the value of that in the interval $(a_1, a_2]$. What is more, the error bound of the 1st degree Taylor expansion is greater than that of the 2nd degree expansion. Therefore, we explore the 2nd degree Taylor expansion to approximate $P_{sm}$ in the interval $[\omega_{sm}^b, a_2]$. Fig. 2(b) shows an example of the 2nd degree Taylor expansion of $P_{sm}$. Given a local point $\omega_{m}^{(k,q)}$ at the $q$-th iteration, the Taylor expansion of $-P_{m}^{(k)}$ at the local point can be given by

$$-\tilde{P}_{m}^{(k)} \approx -P_{m}^{(k)} - \nabla P_{m}^{(k)} (\omega_{m} - \omega_{m}^{(k,q)})^T - \frac{1}{2}(\omega_{m} - \omega_{m}^{(k,q)})H(\omega_{m}^{(k,q)})(\omega_{m} - \omega_{m}^{(k,q)})^T - R_2(\omega_{m})$$

(33)

where $\omega_{m} = [\omega_{1m}, \ldots, \omega_{|S_m|}]^T$, $\nabla P_{m}^{(k)}(\omega_{m})$ is the gradient of $P_{m}^{(k)}$ over $\omega_{m}$ at the local point $\omega_{m}^{(k,q)}$ with

$$\frac{\partial P_{m}^{(k)}}{\partial \omega_{m}^{(k,q)}} = \lambda_1(1+\omega_{m}^{(k,q)})^{-\tau_{s}2} \times \frac{\sum \in \mathcal{S} \Lambda^{1} \left(3.5y_{m}^{2}\omega_{m}^{2.5(k,q)} \right) \left(9y_{m}(\omega_{m} + \omega_{m}^{(k,q)})\right)}{(y_{m} + \omega_{m}^{(k,q)})}$$

(34)

and $H(\omega_{m}^{(k,q)})$ is a Hessian matrix with

$$\frac{\partial^2 P_{m}^{(k)}}{\partial \omega_{m}^{(k,q)} \partial \omega_{m}^{(k,q)}} = \lambda_1(1+\omega_{m}^{(k,q)})^{-\tau_{s}2} \times \frac{10.75y_{m}^{2}\omega_{m}^{2.5(k,q)} \left(9y_{m}(\omega_{m} + \omega_{m}^{(k,q)})\right)}{(y_{m} + \omega_{m}^{(k,q)})} + \frac{\tau_{s}2}{2} \left(9y_{m}(\omega_{m} + \omega_{m}^{(k,q)})\right)$$

(35)

$$\frac{\partial^2 P_{m}^{(k)}}{\partial \omega_{m}^{(k,q)} \partial \omega_{m}^{(k,q)}} = 0, \forall s \neq s'$$

(36)

$y_{sm} = \alpha_{P_{sm}} P_{sm}^{1/2}\rho_{s}^b, \ z_{s} = \alpha_{P_{sm}}$, $\omega_{sm}^{2.5(k,q)}$ rather than $\omega_{sm}^{2.5(k,q)}$ for lightening the notation.

**Lemma 4.** Let the function $\tilde{P}_{m}^{(k)} : \mathbb{R}^{|S_m|} \rightarrow \mathbb{R}$ be three times differentiable in a given interval $[\omega_{sm}^{b}, S_{sm}^*]$ for all $s \in S^I$, then the error bound of the 2nd degree Taylor expansion of $\tilde{P}_{m}^{(k)}$ at the local point $\omega_{sm}^{(k,q)}$ with $\omega_{sm}^{(k,q)} \in [\omega_{sm}^{b}, S_{sm}^*]$ is given by

$$R_2(\omega_{m}) = \frac{1}{3!} \sum_{s \in S^I} (\omega_{sm} - \omega_{sm}^{(k,q)}) \frac{\partial}{\partial \omega_{sm}^{(k,q)}} \tilde{P}_{m}^{(k)} \left|_{\omega_{sm}} \right. \omega_{sm}^{(k,q)}$$

(37)

where $\tilde{P}_{m}^{(k)} \left|_{\omega_{sm}} \right. \omega_{sm}^{(k,q)} = \max \left\{ \tilde{P}_{m}^{(k)} \left|_{\omega_{sm}} \right. \omega_{sm}^{(k,q)}, \tilde{P}_{m}^{(k)} \left|_{S_{sm}^*} \right. \omega_{sm}^{(k,q)} \right\}$, $\omega_{sm}^{b} = [\omega_{sm}^{b1}, \ldots, \omega_{sm}^{b|S_m^I|}]$ and $S_{sm}^* = [S_{sm}^1, \ldots, S_{sm}^{|S_m^I|}]$.

Proof. Please refer to Appendix C.

After conducting the 2nd degree Taylor approximation, the objective function becomes a convex function. Although the constraint (26(b)) is $P_{sm}$ related, we need not to conduct the Taylor approximation on (26(b)) as $P_{sm}$ is quasi-convex and unimodal. In fact, the probability constraint (26(b)) is equivalent to the following inequality

$$\omega_{sm}^{ub} \leq \omega_{sm} \leq \omega_{sm}^{lb}, \forall s \in S^I$$

(38)

where $\omega_{sm}^{ub} \leq W$ represents the upper bound of $\omega_{sm}$ satisfying (26(b)).

Next, a low-complexity bisection-search-based scheme, the main procedures of which are described below, is developed to obtain $\omega_{sm}^{lb}, S_{sm}^*$, and $\omega_{sm}^{ub}$:

a) let the function $Q_{sm} = P_{sm} - \pi_s$. Perform the bisection search method [40] on $Q_{sm} = 0$ to obtain $\omega_{sm}^{lb}$ and $\omega_{sm}^{ub}$ that are the two zero points of $Q_{sm}$,

b) with the obtained $\omega_{sm}^{lb}$ and $\omega_{sm}^{ub}$, find the maximum value $S_{sm}^*$ of $P_{sm}$ using the bisection search method again.

According to the above analysis, at the $q$-th iteration, we can rewrite (32) as

$$\begin{align}
\omega_{sm}^{(k+1,q+1)} = \arg \min_{\omega_{sm}} \tilde{L}^{(q)}(\omega_{sm}^{(k,q)})) \in \arg \min_{\omega_{sm}, G_{i,sm}} \tilde{L}^{(q)}(\omega_{sm}, G_{i,sm})
\end{align}$$

(39a)

subject to:

$$m_{i,sm} = (26(c)) - (26(e)), (38)$$

(39b)

where

$$\tilde{L}^{(q)}(\omega_{sm}, G_{i,sm}) = -\frac{1}{3!} \tilde{P}_{m}^{(k)} - \frac{\tilde{P}(k)}{M} + \sum_{s \in S^I} \left( \psi_{sm}(q) \left( \omega_{sm} - \omega_{sm}^{(k,q)}(I) \right) + \frac{\tau_{s}2}{2} \| \omega_{sm} - \omega_{sm}^{(k,q)}(I) \|^2 \right)$$

In [49], the objective function is convex, (26(c)) is affine, and the constraint (26(d)) can be proved to be convex w.r.t. both $\omega_{sm}$ and $G_{i,sm}$ [22]. Therefore, (39b) is a convex problem that can be effectively mitigated by some standard convex optimization tools such as CVX [2] and MOSEK [7].

Then we can summarize the main steps of mitigating the problem (26) in Algorithm 1.
Algorithm 1 ADMM-based bandwidth allocation algorithm

1: Initialization: Randomly initialize $G_{i,s}^{(0,0)}, \{\omega_{i,s}^{(0,0)}\}$, let $k_{\text{max}} = 250$, $q_{\text{max}} = 250$, $q = 0$, $k = 0$, and generate channel samples $\{H_{i,sm}\}$.
2: repeat
3:   repeat
4:     Given $G_{i,s}^{(k,q)}, \omega_{i,s}^{(k,q)}$, call the greedy scheme to obtain $b_{i,s}^{(k,q)}$.
5:     Optimize (39) with obtained $b_{i,s}^{(k,q+1)}$ to achieve $G_{i,s}^{(k,q+1)}$ and $\omega_{i,s}^{(k,q+1)}$.
6:   Update $q = q + 1$.
7: until Convergence or reach at the maximum iteration times $q_{\text{max}}$.
8: Let $\omega_{i,s}^{(k+1,q+1)} = \omega_{i,s}^{(k,q+1)}$, update $\omega_{i,s}^{(k+1)}$. (40).
9: Call (28) to update $\omega_{i,s}^{(k+1)}(t)$.
10: Update $k = k + 1$.
11: until Convergence or reach at the maximum iteration times $k_{\text{max}}$.

satisfies the low-rank constraint, i.e., the SDR for the power matrix utilized in Algorithm 1 is right.

Proof. Please refer to Appendix D.

VI. OPTIMIZATION OF BEAMFORMING WITH SYSTEM SENSED CHANNELS

In Section V, we obtained a family of global consensus variables $\{\omega_{i}(t)\}$ with the system generated channel samples. The time-varying actual channels may require the re-optimization of beamformers and device associations at each minislot. According to system sensed channels at each minislot, we next discuss how to calculate beamforms and device associations.

At each minislot $t$, given the global consensus variables $\{\omega_{s}(t)\}$, the original problem (24) will be reduced to the following problem

\[
\begin{align*}
\max_{b_{i,s}^{(k)}, G_{i,s}(t)} & \quad \hat{p} U^{(k)}(t) \\
\text{subject to:} & \quad \text{constraints (13), (20), (21)} \text{ are satisfied.}
\end{align*}
\]

In (40), the channels are system sensed ones at $t$. According to the convexity analysis in Section V, (40) is a mixed-integer non-convex programming problem with positive semidefinite matrices, which is hard to be mitigated. Therefore, the alternative optimization scheme presented in subsection V-B can be leveraged to achieve the solutions $b_{i,s}^{(k)}(t)$ and $G_{i,s}(t)$ of (40). Lemma 5 indicates that the achieved rank$(G_{i,s}(t)) \leq 1$. Thus, we can obtain the beamformers $g_{i,s}(t)$ by performing the eigendecomposition on $G_{i,s}(t)$. To sum up, over a time slot $t$, the slice resource optimization algorithm designed for the RAN slicing system can be summarized as follows.

VII. SIMULATION RESULTS

In this section, we aim at evaluating the proposed algorithm via extensive simulations.

Algorithm 2 slice resource optimization algorithm, SRO

1: Initialization: $\{H_{i,s}(t)\}$, $\forall i \in T^u$, $s \in S^u$, and let $P_{s}^{1} \in [0, 1]$, $\mu_{s} = 0$, $\forall s \in S$. $T^u$.
2: Call Algorithm [1] to obtain $\{\omega_{s}(t)\}$ for all $s \in S^t$.
3: for $t = 1 : T$ do
4:   Given $\{\omega_{s}(t)\}$, mitigate (40) by exploiting the alternative optimization scheme to obtain beamformers $\{g_{i,s}(t)\}$ and URLLC device associations $b_{i,s}(t)$ for all $i \in T^u$, $s \in S^u$.
5: end for

| Table II |
|-----------|
| Para. | Value | Para. | Value | Para. | Value |
| $J$ | 3 | $K$ | 2 | $\rho$ | 1 |
| $\eta$ | 100 | $T$ | 60 | $W$ | 60 MHz |
| $M$ | 100 | $\phi$ | 1.5 | $a$ | 0.18 MHz |
| $\xi$ | 54 | $\alpha_9$ | 0.05 | $\kappa$ | 5.12 $\times 10^{-4}$ |
| $\alpha$ | $10^{-5}$ | $\beta$ | $2 \times 10^{-8}$ | $\varsigma$ | $2 \times 10^{-5}$ |

A. Comparison algorithms and parameter setting

We compare the following three algorithms to verify the effectiveness of the proposed algorithm and to explain the impact of access control schemes on the RAN system performance intuitively i) SRO algorithm that adopts the unrestricted access control scheme; ii) SRO-ACB1 algorithm that utilizes the ACB access control scheme with $P_{ACB} = 0.9$; iii) SRO-ACB11 algorithm that adopts the ACB access control scheme with $P_{ACB} = 0.5$.

The parameter setting is as follows: RRHs and IoT devices are deployed following independent PPPs in a one km$^2$ area. URLLC devices are randomly and uniformly distributed in this area. There are three mIoT slices and two URLLC slices in the RAN slicing system. For the mIoT slices, set the new endogenous packet arrivals rate $\lambda_{w,s}(t) = \{1.5,1.0,0.5\}$ packets/minislot, $\pi_s = 0.5$, $\forall s \in T$. Let the path-loss component $\phi = 4$, $L = 2000$ bits, $\tau = 1$ unit, $\sigma^2 = -90$ dBm, $\rho_o = -90$ dBm, $E_i^L = 0.03$ mW, $\lambda_R = 3$ RHs/km$^2$, $\{\lambda_s^R\} = \{18000, 18000, 18000\}$ IoT devices/km$^2$, $\{\gamma_s^L\} = \{5.8, 4.35, 2.9\}$ Kbits/minislot. For the URLLC slices, the transmit antenna gain at each RRH is set to be 5 dB, and a log-normal shadowing path-loss model is leveraged to simulate the path-loss between an RRH and a URLLC device with the log-normal shadowing parameter being 10 dB. A path-loss is computed by $h(\text{dB}) = 128.1 + 37.6 \log_{10} d$, where $d$ (in km) represents the distance between a device and an RRH. Let $L_i^u = 160$ bits, $\sigma_i^L = -100$ dBm, $\lambda_s = \lambda = 0.1$ packets/minislot, $\forall i, s, \{I_s^u\} = \{3, 5\}$ devices, and $\{D_s\} = \{1, 2\}$ milliseconds, $E_j = 3$ W, $\forall j$.

Other simulation parameters are shown in Table II.

B. Performance evaluation

To evaluate the comparison algorithms, the following performance indicators are utilized i) RA success probability $P_s(t)$ that is computed using (9); ii) expected queue length per IoT
device at minislot t, \( E[Q_s(t)] = \mu_{a,s}(t) \); iii) total slice utility \( \bar{U} \) that is the objective function of (24).

We first evaluate the convergence of the proposed SRO algorithm. Fig. 3 illustrates the convergence of SRO with \( \Delta_\omega = \sum_{s \in S^T} \omega^{(k+1)}_s(\bar{I}) - \omega^{(k)}_s(\bar{I}) \). It shows that SRO can converge after several iterations.

We next plot the trend of the RA success probability \( P_s(t) \) and the corresponding expected queue length \( E[Q_s(t)] \) during a time slot in Fig. 4. Fig. 4(a) and 4(c) show the tendency of \( P_s(t) \) and \( E[Q_s(t)] \) in the case of \( \{\gamma_{th}^s\} = \{1.8, 1.35, 0.9\} \). Fig. 4(b) and 4(d) depict the trend of \( P_s(t) \) and \( E[Q_s(t)] \) in the case \( \{\gamma_{th}^s\} = \{5.8, 4.35, 2.9\} \).

From Fig. 4 we obtain the following interesting conclusions: the queue of each IoT device is not stable when the queue serving rate \( \gamma_{th}^s \) is small. In this case, the average queue length monotonically increases over minislots. On the contrary, the queue of each IoT device is periodically flushed when a great queue serving rate is configured.

Let the IoT device intensity \( \lambda^T = [900n, 900n, 900n] \) with \( n \in \{6, 8, \ldots, 26\} \). Under the existence of both mIoT and URLLC slices, we plot trends of the total slice utility \( \bar{U} \) and bursty URLLC slice utility \( \bar{U}^u \) w.r.t. \( n \) in Fig. 5 to understand the impact of the mIoT slices on the performance of all comparison algorithms. In this figure, \( B = [b_{11}^s, b_{31}^s, b_{12}^s, \ldots, b_{52}^s], \omega^T = [\omega^T_{SRO}, \omega^T_{ACB1}, \omega^T_{ACB11}] \) MHz with \( \omega^T_{SRO}, \omega^T_{ACB1}, \omega^T_{ACB11} \) representing the bandwidth allocated to mIoT slices by executing SRO, SRO-ACB1, and SRO-ACB11 algorithms, respectively, and \( \bar{U}^T = [\bar{U}^T_{SRO}, \bar{U}^T_{ACB1}, \bar{U}^T_{ACB11}] \) with \( \bar{U}^T_{SRO} \) denoting the achieved mIoT slice utility of SRO.

The following observations can be obtained from Fig. 5: i) when \( n < 16 \), all algorithms almost obtain the same total slice utility, and the obtained utilities are robust to the average number of IoT devices; ii) when \( 16 \leq n \leq 26 \), the conclusion changes. For the SRO algorithm, its achieved \( \bar{U} \) decreases with an increasing \( n \) due to increasing interference. A great \( n \), however, does not cause a significant decrease in the total slice utilities obtained by SRO-ACB1 and SRO-ACB11. Thanks to the exploration of an access control scheme, both SRO-ACB1 and SRO-ACB11 can achieve greater \( \bar{U} \) than SRO. For example, compared with SRO, SRO-ACB11 improves \( \bar{U} \) by 6.65% when \( n = 24 \); iii) when \( n = 26 \), which means that the total average number of IoT devices reaches 70, 200 devices, the RAN slicing system fails to create and manage mIoT slices as the QoS requirements of mIoT slices serving such a massive average number of devices cannot be simultaneously satisfied. In this case, all system resources are allocated to URLLC slices, and the maximum bursty URLLC slice utility is obtained; iv) as mIoT slices and URLLC slices share the system resources, an increasing \( n \) results in a decreasing bursty URLLC slice utility \( \bar{U}^u \); Besides, it is interesting to find that the two access-control-based algorithms may not outperform SRO in terms of obtaining \( \bar{U}^u \). It indicates that URLLC slices do not benefit from access control schemes of mIoT slices when changing \( n \); v) the RAN slicing system can always accommodate the QoS requirements of all URLLC devices.

Next, to understand the impact of URLLC slices on the
The important observation is that the SRO algorithm when obtained $\lambda$ is changed. ACB obtained mIoT slice utilities by reducing the number of interfering IoT devices; iv) the achieve greater by all algorithms decrease with $\lambda$ to the increase of energy consumption; iii) SRO-ACB is better than SRO. For example, when $\lambda$ of the bursty URLLC slice utility. Two algorithms adopting utility obtained by running the SRO algorithm, respectively.

From Fig. 6 we can observe that: i) the obtained $\bar{U}$ of all algorithms decrease with $\lambda$ mainly due to the decrease of the bursty URLLC slice utility. Two algorithms adopting the access control scheme always achieve greater utilities $\bar{U}$ than SRO. For example, when $\lambda = 5$, compared with the SRO algorithm, the obtained $\bar{U}$ of SRO-ACB is increased by 29.41%; ii) for all algorithms, the computed bandwidth for URLLC slices increases with an increasing $\lambda$. However, their obtained URLLC slice utilities $\bar{U}_u$ are reduced owing to the increase of energy consumption; iii) SRO-ACB may achieve greater $\bar{U}$ than SRO-ACB; as a greater $\bar{U}$ is obtained by reducing the number of interfering IoT devices; iv) the obtained mIoT slice utilities $\bar{U}_u$ of SRO-ACB and SRO-ACB are robust to the URLLC packet arrival rate. The obtained $\bar{U}_u$ of SRO decreases with an increasing $\lambda$; v) an important observation is that the $\bar{U}_u$ of the access-control-based SRO-ACB algorithm is 1.65 times that of the SRO algorithm when $\lambda = 5$. It explicitly reflects that mIoT slices can still benefit from access control schemes even though $\lambda$ is changed.

Figs. 5 and 6 illustrate the situation of a given total system bandwidth. We next change the total bandwidth $W$ and plot its impact on the obtained total slice utilities of all algorithms in Fig. 7.

The following conclusions can be obtained from this figure i) when $W = 45$ MHz, the QoS requirements of all IoT devices cannot be simultaneously satisfied. As a result, the total bandwidth is allocated to URLLC slices; ii) when $W$ locates in the range of $(45, 55]$ MHz, the achieved total slice utilities $\bar{U}$ of SRO and SRO-ACB increase with $W$. Owing to the utilization of the access control scheme, SRO-ACB and SRO-ACB obtain higher $\bar{U}$ than SRO. For example, compared with the SRO algorithm, the SRO-ACB algorithm improves the achieved total slice utility by 6.66% when $W = 50$ MHz. iii) when $W > 55$ MHz, all algorithms cannot remarkably improve $\bar{U}$.

At last, we discuss other crucial parameters’ impact on the performance of the comparison algorithms. We reconfigure $\{\gamma_i\}$ of mIoT slices as $\gamma_1 = 3.6m, \gamma_2 = 2.7m$ and $\gamma_3 = 1.8m$ Kbits/minslot with $m \in \{1.5, 1, \ldots, 2.1\}$ and $\{D_s\}$ of URLLC slices as $D_1 = 0.00025d$ second and $D_2 = 0.0005d$ second with $d \in \{2, 3, \ldots, 10\}$. The impact of QoS requirements of network slices on the total slice utility is plotted in Figs. 8 and 9. The impact of energy efficiency coefficient $\eta$ is plotted in Fig. 10. In this figure, we denote the energy consumption of RRHs of all algorithms by $E^u = \{E^u_{\text{SRO}}, E^u_{\text{ACB}}, E^u_{\text{ACB}}\}$ with $E^u_{\text{SRO}} = \sum_{t=1}^{T} \sum_{a \in S^u} \sum_{b \in a} \epsilon(G_{t,s})$. The impact of energy efficiency coefficient $\eta$ is plotted in Fig. 10.
From these figures, the following observations can be achieved: i) the obtained utilities $\bar{U}$ of all algorithms decrease with an increasing $m$. This is because a greater $m$ indicates that the accumulated IoT packets in the queue of each IoT device can be quickly emptied, and then a smaller $P_r(t)$ is obtained; ii) a great $D_s$ will reduce RRHs’ energy consumption. However, it also reduces the URLLC slice gain. Then, it may be hard to conclude the trend of $\bar{U}^u$ w.r.t. $D_s$ as the energy efficiency coefficient $\eta$ significantly affects the value of $\bar{U}^u$; iii) it is also uneasy to conclude the trend of $\bar{U}^u$ w.r.t. $\eta$. An increasing $\eta$ causes a decrease of RRHs’ energy consumption. Yet, the value of $\bar{U}^u$ is determined by the multiplier of $\eta$ and $E^u$; iv) the SRO-ACB_{II} algorithm may perform better than the SRO algorithm. However, the performance of the other access-control-based algorithm, SRO-ACB_{I}, is slightly worse than the SRO algorithm. Besides, it cannot ensure that the $\bar{U}^t$ obtained by the access-control-based algorithms are always higher than that of SRO. At sometimes, access control schemes may drag down the utility of the mIoT service.

To sum up, in the case of service multiplexing, RA control schemes for alleviating signal interference and enhancing mIoT slice utility may be preferred for mIoT slices. However, considering both the CAPEX and the improvement of slice utility, RA control schemes should be carefully designed and employed because some RA control schemes may worsen the mIoT and even the total slice utilities.

VIII. Conclusion

In this paper, we revisited the frame and minislot structure of a RAN slicing system to admit more IoT devices and proposed a queue evolution model to analyze the RACH of a randomly chosen IoT device. Based on the analysis result, we derived the closed-form expression of the RA success probabilities of the device with unrestricted access control scheme and ACB access control scheme. Next, we formulated the RAN slicing for mIoT and bursty URLLC service multiplexing as an optimization problem to optimally orchestrating RAN resources for mIoT slices and URLLC slices, and efficient mechanisms such as SAA and ADMM were exploited to mitigate the optimization problem. Simulation results showed that RA control schemes should be carefully designed and employed in the case of service multiplexing.

APPENDIX

A. Proof of Lemma 1

For the origin RRH, the LT of its aggregate interference from interfering IoT devices in $s \in S^I$ can be derived as

$$L_{I_s}(\bar{\omega}_s) = E[L_s(t)] + \mathbb{E}[-\mathbb{E}[L_s(t)]]$$

$$= E[L_s(t)] + \mathbb{E}[\exp(-\bar{\omega}_s)]$$

$$\leq E[L_s(t)] + \sum_{m,s} P\{m\} \sum_{s} \mathbb{E}[\eta]$$

$$= E[L_s(t)] + \sum_{m,s} P\{m\} \sum_{s} \mathbb{E}[\eta]$$

$$= E[L_s(t)] + \sum_{m,s} P\{m\} \sum_{s} \mathbb{E}[\eta]$$

$$= E[L_s(t)] + \sum_{m,s} P\{m\} \sum_{s} \mathbb{E}[\eta]$$

$$= E[L_s(t)] + \sum_{m,s} P\{m\} \sum_{s} \mathbb{E}[\eta]$$

$$= E[L_s(t)] + \sum_{m,s} P\{m\} \sum_{s} \mathbb{E}[\eta]$$

$$= E[L_s(t)] + \sum_{m,s} P\{m\} \sum_{s} \mathbb{E}[\eta]$$

$$= E[L_s(t)] + \sum_{m,s} P\{m\} \sum_{s} \mathbb{E}[\eta]$$

$$= E[L_s(t)] + \sum_{m,s} P\{m\} \sum_{s} \mathbb{E}[\eta]$$

where $\bar{\omega}_s = \frac{\gamma(h)}{\rho_0}$, $Z_s$ denotes the set of interfering IoT devices in mIoT slice $s$, $X_s$ represents the number of active IoT devices associated with the origin RRH in $s$. According to the conclusion of Lemma 1 in [1], the probability mass function (PMF) $P_X^s$ can be written as

$$P_X^s(X_s = n') = \frac{1}{(1+\bar{\omega}_s)\rho_0}$$

with $\Gamma(\cdot)$ being the gamma function. Besides, in [1], (a) follows from the i.i.d. distribution of $h_m$ and its further independence from the Poisson point process $\Phi^s$; (b) follows from the expectation of a discrete random variable; (c) follows from the LT over $h_m$; (d) follows from the fact that the number of active IoT device in a specific Voronoi cell is one more than the number of active interfering IoT devices in this cell.

From (42), we can deduce that $X_s (s \in S^I)$ is a gamma-Poisson random variable with $X_s \sim \text{gamma} \sim \text{Poisson}(\alpha, 3.5)$ and $\alpha = \frac{\rho_0}{5\lambda_0 F_s}$.
For a gamma Poisson random variable $X_s \sim \text{gamma} (\alpha, \beta)$, the following expression holds: $E[e^{X_s}] = (1 + \alpha - \alpha e)^{-\beta}$. Thus, we can rewrite (41) as (9). This completes the proof.

B. Proof of Lemma 2

As new endogenous packet arrivals in any IoT device at each minislot $t$ is modelled as a Poisson distribution, the departure process of packets can be regarded as an approximated thinning process of new arrivals, where the thinning factor is related to the RA success probability. The number of accumulated packets in the queue of any IoT device can then be approximated as a Poisson distribution with intensity $\mu_{t,a,s}$ after the thinning process in a specific minislot $t$ ($t > 1$) [29].

Thus, we can derive the expression of $\mu_{t,a,s}^t$ ($t > 1$) via combining with the following facts

- **Fact 1**: the accumulated packets during the $t$-1-th minislot will contribute to the accumulated packets at the $m$-th minislot.
- **Fact 2**: the arrival packets during the $t$-1-th minislot will also contribute to the accumulated packets in the queue of an IoT device at the $m$-th minislot.
- **Fact 3**: an IoT device can send packets only if its preamble is successfully transmitted.
- **Fact 4**: at the same minislot, the new packet arrival process and the packet accumulated process are independent.

Similar as the Theorem 2 in [29], we can infer that at the 2nd minislot, for all $s \in S^I$, $\mu_{t,a,s}^2$ depends on the intensity of new packet arrivals $\mu_{t,a,s}^1$ and the probability $P_s^1$ of a randomly selected IoT device at the 1st minislot, which is given by

$$
\mu_{t,a,s}^2 = \mu_{t,a,s}^1 - x_sP_s^1(1 - e^{-\mu_{t,a,s}^1})
$$

The detailed proof of (43) is omitted for brevity, and a similar proof can be found in the proof section of Theorem 2 in [29].

Considering that $\mu_{t,a,s}(t)$ is non-negative at each minislot $t$, we have

$$
\mu_{t,a,s}^2 = \left[ \mu_{t,a,s}^1 - x_sP_s^1(1 - e^{-\mu_{t,a,s}^1}) \right]^{+}
$$

Then, according to the definition of non-empty probability and the Poisson approximation, the non-empty probability of a randomly selected IoT device in mIoT slice $s \in S^I$ at the 2nd minislot can be approximated as

$$
P_{n,s}^2 = 1 - e^{-\mu_{t,a,s}^2}
$$

At the 3rd minislot, the intensity of accumulated data packets in the queue of a randomly selected IoT device can be derived as the following

$$
\mu_{t,a,s}^3 = P_s^2 \left( \sum_{n=1}^{\infty} \left( \frac{\mu_{t,a,s}}{1 - e^{-\mu_{t,a,s}}} \right)^n \right) + (1 - P_s^2) \left( \sum_{n=0}^{\infty} \frac{\mu_{t,a,s}^n}{(n-1)!} \right)
$$

where $P_{N_{s},n}^2$ and $P_{N_{s},n}^2$ represent the PMFs of new arrival packets and accumulated packets at the 2nd minislot, respectively. Besides, (a) follows from the fact: for any two independent Poisson distribution $X_s$ and $X_s$, $P_{X_s}(X_1 + X_2) = \sum_{x=0}^{\infty} P_{X_s}(X_1 = x)P_{X_s}(X_2 = x - y)$; (b) holds as $X_s$ is a two dimensional Poisson distribution with an intensity $\lambda_{X_s}$ and $\lambda_{X_s}$, and $\sum_{x=1}^{\infty} P_{X_s}(X_1 + X_2 = x) = 1 - P_{X_s}(X_1 + X_2 = 0)$.

Similarly, we have

$$
P_{n,s}^3 = 1 - e^{-\mu_{t,a,s}^3}
$$

When $t > 3$, since the accumulated packets evolution model of the queue of any IoT device is similar as that at $t = 3$, we can directly extend the conclusion obtained at $t = 3$ to that at $t > 3$.

Therefore, we can obtain the closed-form expression of $\mu_{t,a,s}^t$ for all $s \in S^I$ at $t > 1$ with

$$
\mu_{t,a,s}^t = \left[ \mu_{t,a,s}^{t-1} - x_sP_s^{t-1}(1 - e^{-\mu_{t,a,s}^{t-1}}) \right]^{+}
$$

and

$$
P_{n,s}^t = 1 - e^{-\mu_{t,a,s}^t}
$$

This completes the proof.

C. Proof of Lemma 4

The 2nd degree Taylor expansion of $\tilde{P}_m^{(k)}$ at the local point $\omega_m^{(k,q)}$ is

$$
\tilde{P}_m^{(k)} = \sum_{j=0}^{2} \frac{1}{j!} \left[ \sum_{x \in S^I} \left( \omega_{sm} - \omega_{sm}^{(k,q)} \right) \frac{\partial}{\partial \omega_{sm}} \right]^j \tilde{P}_m^{(k)} |_{\omega_m^{(k,q)}}
$$

The 3rd degree Taylor expansion of $\tilde{P}_m^{(k)}$ at $\omega_m^{(k,q)}$ must be more accurate than $\tilde{P}_m^{(k)}$ with

$$
\tilde{P}_3^{(k)} = \tilde{P}_m^{(k)} + \frac{1}{3!} \left[ \sum_{x \in S^I} \left( \omega_{sm} - \omega_{sm}^{(k,q)} \right) \frac{\partial}{\partial \omega_{sm}} \right]^3 \tilde{P}_m^{(k)} |_{\omega_m^{(k,q)}}
$$

Since the error of $\tilde{P}_m^{(k)}$ is no greater than the maximum difference between $\tilde{P}_3^{(k)}$ and $\tilde{P}_m^{(k)}$, we have

$$
R_m^3(\omega_m) = \max \left( \frac{1}{3!} \left[ \sum_{s \in S^I} \left( \omega_{sm} - \omega_{sm}^{(k,q)} \right) \frac{\partial}{\partial \omega_{sm}} \right]^3 \tilde{P}_m^{(k)} |_{\omega_m^{(k,q)}} \right)
$$
In (52), $\omega^{(k,q)}_m$ is a constant vector, the max operation will not affect the constant vector and the vector $\omega^{(k,q)}_m$. For any $s \in S^I$, the maximum value obtainable by $\frac{\partial^3 P^i_{\omega^{(k,q)}_m}}{\partial \omega^{(k,q)}_m}$ will not exceed the greatest value of that derivative in the interval $[\omega^{(k,q)}_m, S^*_m]$. Additionally, the maximum value of $\frac{\partial^3 P_m|_{\omega^{(k,q)}_m}}{\partial \omega^{(k,q)}_m}$ will generally occur at one of the endpoints of the interval $[\omega^{(k,q)}_m, S^*_m]$. Therefore, we obtain (37). This completes the proof.

D. Proof of Lemma 5

For all $i \in I^u$, $s \in S^u$, $m \in M$, a feasible way of proving that $\text{rank}(G_{i,s,m}) \leq 1$ is to utilize the Lagrange method. However, owing to the complicated expression of $W^u(r_m)$ w.r.t. $G_{i,s,m}$ it will be uneasy to do that. Fortunately, we find that the proof can be conducted if a family of auxiliary variables is introduced.

For the constraint (264), if we introduce the auxiliary variables $\{\nu_{i,s,m}\}$ and let

$$\frac{\text{tr}(H_{i,s,m}G_{i,s,m})}{\sigma^2_{t,s}} \geq \nu_{i,s,m}, \forall i \in I^u, s \in S^u, m \in M \quad (53)$$

then (264) is equivalent to

$$\sum_{s \in S^I} \left(1 + \alpha_g \omega_{s,m}(i)\right) + W^u(f_m) \leq W, \quad \text{and} \quad (54)$$

where $f_m = \{f_{i,s,m}: i \in I^u, s \in S^u\}$ and

$$f_{i,s,m} = \frac{L^2_i}{\log_2(1 + \nu_{i,s,m})} + \frac{(Q^{-1}(\beta))^2}{2 \log_2(1 + \nu_{i,s,m})}$$

$$+ \frac{(Q^{-1}(\beta))^2}{2 \log_2(1 + \nu_{i,s,m})} \left(1 + 4L^2_i \log_2(1 + \nu_{i,s,m}) \right) (55)$$

We omit the proof of the equivalence as a similar proof can be found in the proof section of constraints’ equivalence in (22).

The partial Lagrangian function of (39) can be written as

$$L(\ldots) = \sum_{s \in S^u} \sum_{i \in I^u} \left[ \frac{\hat{m}}{M} \text{tr}(G_{i,s,m}) - \tilde{\nu}_{i,s,m} \frac{\text{tr}(H_{i,s,m}G_{i,s,m})}{\sigma^2_{t,s}} + \sum_{j \in J} \tilde{\lambda}_{jm} \text{tr}(b^{u(k,q)}_{i,s,m} Z_j G_{i,s,m}) - \tilde{X}_{i,s,m} G_{i,s,m} \right]$$

(56)

where $\tilde{\lambda}_{jm}, \tilde{\nu}_{i,s,m}$, and $\tilde{X}_{i,s,m}$ are Lagrangian multipliers corresponding to constraints (260), (53) and (26b). Besides, only terms related to $G_{i,s,m}$ are included in this function for brevity.

According to the Karush-Kuhn-Tucker (KKT) conditions, the necessary condition for obtaining the optimal matrix power at the $(k,q)$-th iteration $G_{i,s,m}^{(k,q)*}$ is given by

$$-\frac{\partial L(\ldots)}{\partial G_{i,s,m}} = \frac{\hat{m}}{M} I_{i,s,m} + \tilde{\nu}_{i,s,m} H_{i,s,m} - \sum_{j \in J} \tilde{\lambda}_{jm} b^{u(k,q)}_{i,s,m} Z_j - \tilde{X}_{i,s,m} = 0 \quad (57)$$

where $I_{i,s,m} \in \mathbb{R}^{JK \times JK}$ is an identity matrix.

Then, we can conclude that $\text{rank}(X_{i,s,m}) \geq JK - 1$. The reasons are i) $\tilde{\lambda}_{jm}, b^{u(k,q)}_{i,s,m}$, and $\tilde{\nu}_{i,s,m}$ are nonnegative and the matrix $I_{i,s,m}$ is full rank; ii) $\text{rank}(H_{i,s,m}) \leq 1$.

Next, according to the complementary slackness condition, we have

$$X_{i,s,m} G_{i,s,m}^{(k,q)*} = 0 \quad (58)$$

Based on (58) and the rank result of $X_{i,s,m}$, we can conclude that $\text{rank}(G_{i,s,m}^{(k,q)*}) \leq 1$. This completes the proof.

REFERENCES

[1] Ericsson, “Cellular networks for massive IoT,” Ericsson, Stockholm, Sweden, Tech. Rep. Uen 284 23-3278, Jan. 2016.

[2] Nokia. "LTE evolution for IoT connectivity. " Nokia, Espoo, Finland, Espoo, Finland, Tech. Rep. SR1702006775EN, Jan. 2016.

[3] S. Xing, X. Wen, Z. Lu, Q. Pan, and W. Jing, “A novel distributed queuing-based random access protocol for narrowband-IoT,” in ICC 2019-2019 IEEE International Conference on Communications (ICC). IEEE, 2019, pp. 1–7.

[4] H. S. Jang, H.-S. Park, and D. K. Sung, “A non-orthogonal resource allocation scheme in spatial group based random access for cellular M2M communications,” IEEE Transactions on Vehicular Technology, vol. 66, no. 5, pp. 4496–4500, 2016.

[5] N. Jiang, Y. Deng, A. Nallanathan, and J. A. Chambers, “Reinforcement learning for real-time optimization in NB-IoT networks,” IEEE Journal on Selected Areas in Communications, vol. 37, no. 6, pp. 1442–1450, 2019.

[6] Z. Zhang, Y. Li, C. Huang, Q. Guo, C. Yuen, and Y. L. Guan, “DNN-aided block sparse bayesian learning for user activity detection and channel estimation in grant-free non-orthogonal random access,” IEEE Transactions on Vehicular Technology, vol. 68, no. 12, pp. 12000–12012, 2019.

[7] N. Jiang, Y. Deng, A. Nallanathan, X. Kang, and T. Q. Quek, “Analyzing random access collision losses in massive IoT networks,” IEEE Transactions on Wireless Communications, vol. 17, no. 10, pp. 6853–6870, 2018.

[8] M. Grau, C. H. Foh, A. Ul Quddus, and R. Tafazoli, “Preambral barring: A novel random access scheme for machine type communications with unpredictable traffic bursts,” in 2019 IEEE 98th Vehicular Technology Conference (VTC2019-Fall). IEEE, 2019, pp. 1–7.

[9] H. S. Jang, H. Jin, B. C. Jung, and T. Q. Quek, “Recursive access class barring for machine type communications with PUSCH resource constraints,” in ICC 2019-2019 IEEE International Conference on Communications (ICC). IEEE, 2019, pp. 1–6.

[10] H. Jin, W. T. Toor, B. C. Jung, and J.-B. Seo, “Recursive pseudo-bayesian access class barring for M2M communications in LTE systems,” IEEE Transactions on Vehicular Technology, vol. 66, no. 9, pp. 8595–8599, 2017.

[11] H. S. Jang, H. Jin, B. C. Jung, and T. Q. Quek, “Versatile access control for massive IoT: Throughput, latency, and energy efficiency,” IEEE Transactions on Mobile Computing, 2019, in press. DOI 10.1109/TMC.2019.2914381.

[12] N. Allianc, “5G white paper,” Next generation mobile networks, white paper, vol. 1, 2015.

[13] 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, 2018.

[14] I. Bushiraj, S. Tyagi, S. Tarwar, N. Kumar, and J. J. Rodrigues, “Tactile internet for smart communities in 5G: An insight for NOMA-based solutions,” IEEE Transactions on Industrial Informatics, vol. 15, no. 5, pp. 3104–3112, 2019.

[15] A. Ksentini and N. Nikaein, “Toward enforcing network slicing on RAN: Flexibility and resources abstraction,” IEEE Communications Magazine, vol. 55, no. 6, pp. 102–108, 2017.

[16] P. Rost, C. Mannweiler, D. S. Michalopoulos, C. Sartori, V. Sciancalepore, N. Sastry, O. Holland, S. Tayade, B. Han, D. Bega et al., “Network slicing to enable scalability and flexibility in 5G mobile networks,” IEEE Communications magazine, vol. 55, no. 5, pp. 72–79, 2017.

[17] R. Wen, G. Feng, J. Tang, T. Q. Quek, G. Wang, W. Tan, and S. Qin, “On radio resource efficiency of network slicing for next-generation mobile networks,” IEEE Transactions on Communications, vol. 67, no. 1, pp. 430–444, 2018.

[18] T. V. K. Buyakar, A. PC, B. R. Tamma et al., “Poster: Scalable network slicing architecture for 5G,” in Proceedings of the 24th Annual International Conference on Mobile Computing and Networking, ACM, 2018, pp. 684–686.

[19] T. Guo and A. Suárez, “Enabling 5G RAN slicing with EDF slice scheduling,” IEEE Transactions on Vehicular Technology, vol. 68, no. 3, pp. 2865–2877, 2019.
[20] 3GPP, “Cellular system support for ultra low complexity and low throughput Internet of Things,” the third-generation partnership project, Tech. Rep. v. 13.1.0, Nov. 2015, http://www.3gpp.org/ftp/Specs/archive/45_series/45.820/45820-d10.zip.

[21] J. Tang, B. Shim, and T. Q. Quek, “Service multiplexing and revenue maximization in sliced C-RAN incorporated with URLLC and multicast eMBB,” IEEE Journal on Selected Areas in Communications, vol. 37, no. 4, pp. 881–895, 2019.

[22] P. Yang, X. Xi, T. Quek, J. Chen, X. Cao, and O. D. Wu, “How should I orchestrate resources of my slices for bursty URLLC service provision?” Tech. Rep. Online, 2019, https://arxiv.org/pdf/1912.00579.pdf.

[23] M. N. Soorki, W. Saad, M. H. Manshadi, and H. Saldi, “Stochastic coalitional games for cooperative random access in M2M communications,” IEEE Transactions on Wireless Communications, vol. 16, no. 9, pp. 6179–6192, 2017.

[24] M. Gharbieh, H. ElSawy, A. Bader, and M.-S. Alouini, “Spatial-temporal stochastic modeling of IoT enabled cellular networks: Scalability and stability analysis,” IEEE Transactions on Communications, vol. 65, no. 8, pp. 3585–3600, 2017.

[25] S. Kaul, M. Gruteser, V. Rai, and J. Kenney, “Minimizing age of information in vehicular networks,” in 2011 8th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks. IEEE, 2011, pp. 350–358.

[26] S. Kaul, R. Yates, and M. Gruteser, “Real-time status: How often should one update?” in 2012 Proceedings IEEE INFOCOM. IEEE, 2012, pp. 2731–2735.

[27] 3GPP, “Study on RAN improvements for machine-type communications,” 3GPP, Sophia, Antipolis, France, Tech. Rep. TR 37.868 V11.0.0, Sep. 2011.

[28] N. Jiang, Y. Deng, A. Nallanathan, X. Kang, and T. Q. Quek, “Collision analysis of mIoT network with power ramping scheme,” in 2018 IEEE International Conference on Communications (ICC). IEEE, 2018, pp. 1–7.

[29] N. Jiang, Y. Deng, X. Kang, and A. Nallanathan, “Random access analysis for massive IoT networks under a new spatio-temporal model: A stochastic geometry approach,” IEEE Transactions on Communications, vol. 66, no. 11, pp. 5788–5803, 2018.

[30] M. Haenggi, Stochastic geometry for wireless networks. Cambridge University Press, 2012.

[31] H. ElSawy and E. Hossain, “On stochastic geometry modeling of cellular uplink transmission with truncated channel inverse power control,” IEEE Transactions on Wireless Communications, vol. 13, no. 8, pp. 4454–4469, 2014.

[32] H. Zhang, C. Jiang, N. C. Beaulieu, X. Chu, X. Wang, and T. Q. Quek, “Resource allocation for cognitive small cell networks: A cooperative bargaining game theoretic approach,” IEEE Transactions on Wireless Communications, vol. 14, no. 6, pp. 3481–3493, 2015.

[33] A. Anand and G. de Veciana, “Resource allocation and HARQ optimization for URLLC traffic in 5G wireless networks,” IEEE Journal on Selected Areas on Communications, vol. 36, no. 11, pp. 2411–2421, 2018.

[34] Z. Hou, C. She, Y. Li, T. Q. Quek, and B. Vucetic, “Burstiness-aware bandwidth reservation for ultra-reliable and low-latency communications in tactile internet,” IEEE Journal on Selected Areas in Communications, vol. 36, no. 11, pp. 2401–2410, 2018.

[35] W. Yuen, G. Durisi, T. Koch, and Y. Polyanskiy, “Quasi-static multiple-antenna fading channels at finite blocklength,” IEEE Transactions on Information Theory, vol. 60, no. 7, pp. 4232–4265, 2014.

[36] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein et al., “Distributed optimization and statistical learning via the alternating direction method of multipliers,” Foundations and Trends® in Machine learning, vol. 3, no. 1, pp. 1–122, 2011.

[37] W.-K. K. Ma, “Semidefinite relaxation of quadratic optimization problems and applications,” IEEE Signal Processing Magazine, vol. 1053, no. 5888/10, 2010.

[38] P. Yang, X. Cao, X. Xi, Z. Xiao, and D. Wu, “Three-dimensional drone-cell deployment for congestion mitigation in cellular networks,” IEEE Transactions on Vehicular Technology, vol. 67, no. 10, pp. 9867–9881, 2018.