IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 16, 2020, accepted February 8, 2020, date of publication February 18, 2020, date of current version February 27, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2974812

Dynamic Resource Allocation With RAN Slicing and Scheduling for uRLLC and eMBB Hybrid Services

LEI FENG[1], **YUEQI ZI**[1], **WENJING LI**[2], **FANQING ZHOU**[1], **PENG YU**[1], **AND MICHEL KADOCH**[3]

[1]Institute of Network Technology, Beijing University Posts and Telecommunications, Beijing 100876, China
[2]State Key Laboratory of Networking and Switching Technology, Beijing University Posts and Telecommunications, Beijing 100876, China
[3]Department of Electrical Engineering, Ecole de technologie superieure, Montreal, QC H3C 1K3, Canada

Corresponding author: Wenjing Li (wjli@bupt.edu.cn)

This work was supported in part by the Beijing Natural Science Fund-Haidian Original Innovation Joint Fund under Grant L192003, and in part by the Fundamental Research Funds for the Central Universities through the State Grid Science and Technology project "Analysis of Power Wireless Private Network Evolution and 4G/5G Technology Application" under Grant 2019RC09.

ABSTRACT To cope with the limited radio and power resources, designing energy- and cost-efficient resource allocation strategy with RAN slicing and scheduling is important in ensuring the extreme QoS of differentiated Internet of things (IoT) services. In this regard, we focus on guaranteeing the latency and reliability of sporadic uRLLC uplink traffic while improving the quality of continuous eMBB services (e.g., quality of the video) together in this paper. Firstly, a dynamic optimization model considering power consumption and service quality is used to construct the cost function in both time domain and frequency bandwidth for heterogeneous services, subject to the latency constraint. Secondly, given its complexity, a novel two-timescale algorithm with employing Lyapunov optimization is designed, including two sub-algorithms: long-timescale bandwidth allocation and short-timescale service control. In further, the theoretical optimality is analyzed according to the relationships between control parameters and service performances. The utility of our approach and its hard latency guarantee are also illustrated through simulation results under tolerable power consumption.

INDEX TERMS 5G, network slicing, radio access network, resource allocation, heterogeneous services.

I. INTRODUCTION

As the evolution of wireless technologies, one important motivation of the future cellular networks is to support heterogeneous services which have distinctive and extreme requirements for network performance. Especially, the devices of Internet of things (IoT) require power reduction. And ultra-reliable low latency communication (uRLLC) requires extremely low latency (0.25-0.3 ms/packet) and high reliability (99.999%) [1] and enhanced mobile broadband (eMBB) requires high bandwidth (several 100 MHz to support gigabit per second for high peak data rates) [2]. Since the traditional one-size-fits-all network infrastructure may not accommodate a large expansion of heterogeneous services simultaneously, network slicing (NS) comes into play. With NS, the radio network resource can be partitioned into logically independent resources, called radio access network (RAN) slices, to provide a service-specific RAN [3]. Each slice is customized and orchestrated to only provide what is necessary for one kind of service, avoiding unnecessary overheads and complexity [4]. Then heterogeneous services can be aggregated at access sides and supported by a unified physical infrastructure.

Due to the scarcity of inherent radio resources, a surge in network traffic volume and densification of devices pose technical challenges on implementing RAN slicing and scheduling for differentiated services. Take the typical two mainstream services, uRLLC and eMBB, as an example. A video streaming eMBB traffic in one cell would like a sufficiently high and stable quality of image or voice contents so that some radio resources need to be guaranteed over its transmission time-interval [5]. While the power reduction and ultra-low delay are common goals of IoT devices for uRLLC services. If the uRLLC traffic is activated by some critical

The associate editor coordinating the review of this manuscript and approving it for publication was Adnan Shahid.

outside events in the same cell, it will rapidly occupy these radio resources to obtain a low-latency performance and save energy consumption as much as possible for future critical communication. However, filling the corresponding latency requirement of uRLLC comes at the expense of a degraded QoS (quality of service) of eMBB and network utility.

In recent years, multi-operator RAN sharing based on virtualization has been studied extensively [6], [7]. However, these studies only target on achieving optimal network utility or resource utilization and ignore an explicit characterization underlying extreme service requirements for heterogeneous services. Some existing studies [1], [8] propose an innovative super-position/puncturing framework for multiplexing uRLLC and eMBB traffic. Since the principle is to overwrite part of ongoing eMBB transmissions when sporadic uRLLC traffic appears, the approach could result in QoS deterioration of eMBB services.

On the other hand, the realistic channel condition may be frequency selective fading. RAN resource also needs to be dynamically sliced and scheduled according to the preference of heterogeneous services and future information (e.g., service demands and channel states) in order to keep or update the assignment in case of traffic or service changes. However, such knowledge is difficult to predict because of time-varying characteristic and so on. In this regard, 3GPP has proposed an statically approach based on "fixed network resource shares" [9]. However, the problem statement must not be limited to static wireless resource scheduling as in [10], [11], but also has to consider the instantaneous channel quality and diversity in demands, in order to support such versatile and ambitious use cases.

The aforementioned two respects raise extreme challenges for dynamical RAN slicing and scheduling framework to accommodate hybrid services. In this regard, this paper proposes a dynamic resource allocation scheme, aiming at jointly optimizing the power consumption and bandwidth allocation while satisfying the corresponding latency for sporadic uRLLC traffic arrivals and the quality of eMBB services as much as possible. Bandwidth allocation guarantees the quality of eMBB services, which modeled by maximizing the network utility with assigning a private share. The private share stands for the priority of each slice which employs fair utility for better resource allocation according to the service requirements, thereby asymptotically maximizes utility for eMBB users. The main contributions of this paper can be summarized as follows.

- Previous work only focuses on achieving optimal network utility or resource utilization without considering explicit characterization underlying extreme service requirements. Therefore, we discuss the extreme service requirements. The resource allocation model is constructed to describe the energy-efficiency requirement of mission-critical IoT devices, network utility guarantee and latency constraints. To establish their relationships, we employ a queue update model with queue backlog, generated data and transmission rate which also extra considers reliability. Furthermore, the design of dynamic framework for bandwidth allocation, service controlling and scheduling are also discussed.
- Concluded from previous research, using a super-position/puncturing framework for multiplexing uRLLC and eMBB traffic will result in QoS deterioration of eMBB services. In order to adapt to the time-varying information and solve this problem, we extend the Lyapunov optimization approach to two different time scales aimed at facilitating the cost vs. latency trade-off under reliability constraint. A novel two timescale resource allocation strategy is also designed with deriving analytical bounds of the proposed problem.
- Existing studies mainly talk about network-level spectrum sharing without explicit characterization for heterogeneous services. Thus network slicing has been proposed as a promising paradigm to solve this problem. Given its computational complexity, we address the problem with an alternative minimization algorithm including two sub-algorithms: the latency and transmit power control (LTPC) for uRLLC services and service quality decision (SQD) for eMBB services using network slicing technology. The proposed two sub-algorithms can be guaranteed to converge to the global optimal solutions by Gauss-Seidel method. Heterogeneous services with limited resources now satisfy both uRLLC latency and eMBB throughput requirements.

The numerical simulation results verify that our algorithm outperforms BSRA (Bandwidth Slicing and Resource Allocation algorithm) [12], ACS (Alternative Concave Search algorithm) [13] and Loading Balancing [14] ones in terms of hard latency and power consumption. Moreover, the proposed algorithm can implement a similar performance compared with the semi-offline method which ideally assumes that the future IoT service patterns and network channel information can be learned in advance.

The structure of this paper is as follows. In Section II, we present relevant work, outlining an overview of existing contributions and shortcomings in regard to network resource sharing, bandwidth scheduling and active or passive dynamic resource allocation. In Section III, we formally propose the system model with introducing the transmission rate, network utility and latency of queue. Section IV formulates the sliced RAN uplink resource allocation problem, introduces the Lyapunov optimization, and designs the two timescale and alternative minimization algorithms. Section V evaluates the performance of proposed algorithms and analyzes results. Finally, the paper summarizes our findings in Section VI.

II. RELATED WORK

The implementation of the slice resource allocation is essentially similar to the principle of radio resource sharing in wireless communication networks. In recent years, research on multi-operator RAN sharing based on virtualization has received great progress. Among them, Network Function

Virtualization (NFV), as a promising enabling technology, can effectively promote spectrum sharing between heterogeneous service devices without increasing network deployment costs [15], [16]. For instance, reference [17] compares the sharing between two kinds of resources, physical resource sharing and virtual resource sharing. The results show that virtual resource sharing provides better performance at the expense of complexity. For RAN resource sharing, reference [18] speculates on congestion of future request traffic and uses a random auction approach to solve the capacity allocation problem between multiple virtual network operators (VNO). Reference [19] introduces a dynamic queue to analyze the average delay of data which is generated by mobile terminals in downlink transmission and proposes a delay-aware resource sharing scheme for RAN. Considering that network slice is a collection of similar services, a larger number of network resource sharing methods provide some valuable algorithms that can be used for allocating resources to slices in wireless networks. However, existing studies mainly focus on network-level spectrum sharing under the traditional cellular architecture without an explicit characterization on heterogeneous services, which will not be able to fulfill the requirements of current fast-growing broadband multimedia services and ultra-low latency services.

In response to the above issues, network slicing technology has drawn extensive attention due to its advantages in supporting a surge of hybrid services and improving resource utilization in RAN [13], [20]. Reference [21] formulates the NS framework as a weighted throughput maximization problem for multi-tenant H-CRANs (heterogeneous cloud radio access networks) which defines NS as a process of sharing computational resources. On the other hand, there are also many algorithms for the multi-tenant RAN slicing [7], [22]. By assigning different shares for operators in the network infrastructure, the operators with higher share are expected to get more resources with a higher priority. Specifically, both two algorithms have proposed the static resource allocation which can customize slices, aiming at maximizing network utility. Although the above research mentioned network slicing, none of them considers the heterogeneous services and their extreme performance requirements in terms of hard latency.

In the context of the existing resource allocation models described above, this work covers the following gap in the literature: a dynamic resource allocation scheme for heterogeneous services with limited resources, meeting both uRLLC latency and eMBB throughput requirements. Reference [23] considers the service performance factor (i.e., delay) and studies the bandwidth allocation strategy for edge computing traffic offloading. However, 5G future networks need to carry numerous heterogeneous services, generally classified as uRLLC, eMBB and mMTC (massive machine-type communications). Compared to optimize latency and power consumption, some eMBB services, such as video streaming, prefer high throughput to guarantee the sufficiently high and stable quality of image or voice contents. Delay-optimized resource allocation may result in excessive consumption of network resources due to unnecessary latency reduction. In addition, from a more practical angle, the strategy for optimizing the energy consumption generated in data transmission at the expense of increased service delay is of less significance. The main reason is that the average mobile device generates less energy in sending requests, which is far lower than the power generated by the central processing unit (CPU).

In this paper, we focus specially on the design of dynamic framework for uRLLC and eMBB which has been previously investigated by [1], [8]; however, all these works differ substantially from ours in terms of scope, criterion or approach. Reference [1] proposes an innovative superposition/puncturing framework for multiplexing uRLLC and eMBB traffic in 5G cellular systems, taking into account various models for the eMBB rate loss associated with uRLLC superposition/puncturing. Furthermore, reference [8] develops a joint multi-user preemptive scheduling strategy to simultaneously cross-optimize system spectral efficiency and uRLLC latency. The principle is that the uRLLC traffic instantly overwrites part of the ongoing eMBB transmissions when the network appears sporadic uRLLC traffic, which considers a fundamentally different problem from the one addressed in our paper.

Dealing with dynamic resource allocation, two algorithms are widely studied: active and passive resource allocation. Reference [24] proposes an active pre-allocation mechanism for virtual radio resources, including inter-chip preallocation and on-chip scheduling, but it is limited by the accuracy of the model. When the prediction error occurs, resource redundancy or deficiency would be touched off, resulting in destructive QoS. Reference [25] proposes a time-dependent pricing bandwidth consumption scheme for multimedia streaming applications. The principle is that software defined network (SDN) technology allowed multimedia streaming media users to negotiate their QoS parameters as needed, encouraging delay-tolerant users to release resources for latency-sensitive services in advance. Indeed, due to the unpredictable characteristics of future service requirements and channel state, the solution is not effective in dealing with sudden heterogeneous service requirements, especially the uRLLC service. In order to shield time-varying network conditions, passive resource allocation is more widely applied to radio resource scheduling. A common approach is to directly process the current load task with the queue, allocating resources accurately without future wireless channel states. A related dynamic resource allocation model often considered for heterogeneous services is the so-called "Bandwidth Slicing and Resource Allocation (BSRA) algorithm" [12]. The algorithm proposes an algorithm for power allocation and quality decision in a coexistence scenario of uplink IoT services and downstream video streaming services. However, the scheme has been limited to the case of resource allocation for delay-tolerant services, which can lead to poor performance in the heterogeneous services with uRLLC requests.

In the following, we present a dynamic resource allocation framework to facilitate RAN slicing among heterogeneous services. It aims to jointly optimize the bandwidth allocation and power consumption for IoT devices while satisfying the corresponding latency for sporadic uRLLC traffic arrivals and the quality of eMBB services as much as possible. To describe the dynamic model, we extend Lyapunov optimization to a novel two timescale scheme to solve the proposed problem. An alternative minimization algorithm is then designed to obtain the optimal resource allocation decisions in theory according to the stabilized threshold. The numerical results verify the effectiveness of our algorithm.

III. SYSTEM MODEL

In this section, we describe the mathematical model of resource allocation including RAN and latency of queue update.

A. TRANSMISSION RATE FOR HYBRID SERVICES

A coexistence scenario of two different wireless services (i.e. uRLLC service and eMBB service) is considered in SDN-enabled wireless RAN to allocate the shared resource. The uRLLC traffic is characterized by small and sporadically data-package, requiring ultra-low latency and high reliability. 3GPP TR38.913 regulates uRLLC's latency indicators that uplink and downlink latency of its user plane are also limited to 0.5ms. On the contrary, eMBB services are featured with large payloads, considering high peak data rates or other high bandwidth to guarantee sufficiently high and stable quality (e.g., of image or voice contents). To support such services simultaneously, the controller allocates corresponding resources for specific network slices and controls the performance of IoT devices.

The details of dynamic resource allocation framework are shown in Fig. 1. The concept of queue backlog is used to describe the relationship between the amount of generated data and transmitted data, which updates over time. During the whole RAN transmission process, the generate traffic such as traffic control and video information are transferred to/from data centers in the core networks through a BS (Base station). The power of IoT devices and the quality of video then can be controlled in response to the network condition [12]. We reserve some radio resources for uRLLC and eMBB slices in advance, while sharing others according to the queue backlog to customize the frequency bandwidth slice.

According to the characteristics of two heterogeneous services, we propose a dynamic algorithm with two timescales. The long timescale provides dynamic radio resource allocation policies for two different services, subject to the latency deadline. The short timescale dynamically controls the power consumption of IoT devices and the quality of eMBB services. To model the time division system, we use t to represent a time-slot and $T(= mt, m = 0, 1, 2, \ldots)$ to represent the long timescale, i.e., a time period. At each time slot $t_k(= kT, k = 0, 1, 2, \ldots)$, the system performs the update of bandwidth slicing and allocation strategy. For each time slot $t \in [t_k, t_k + T - 1]$, the system will control and schedule the service request of the hybrid service, i.e., the power control and quality decision. In addition, the dynamic arrival and transmission of requests are represented by queue updating, which also used to describe the latency.

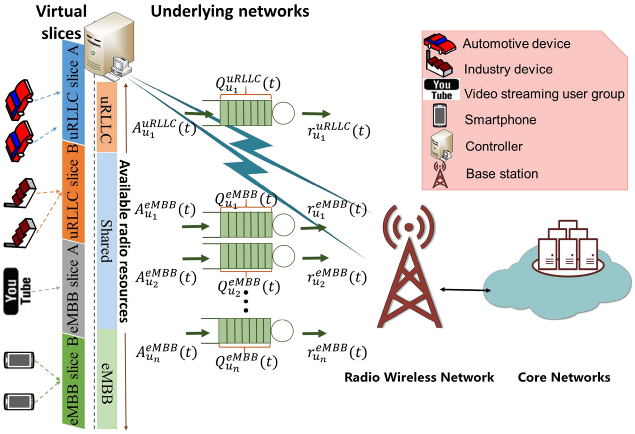


FIGURE 1. Radio resource allocation framework for Network slicing.

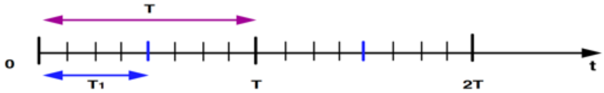


FIGURE 2. An example of different time scales T and T_1.

In this system, time is divided into long timeslots which are further sub-divided into short timeslots. From a scheduling perspective, eMBB and uRLLC resource allocations occur at long timeslot and service control occurs at the short timeslot. Long timeslot length of our model scales with the sub-carrier spacing, where sub-carrier spacing is 30kHz and slots per sub-frame is 2, slot length is 1 ms/2 (=500μs). The short timescale is subject to the granularity of mini-slot, dynamically controlling the energy consumption of IoT devices and the request quality of eMBB services. An example of these different time scales is given in Fig. 2. In this example, $T(=8)$ denotes long timescale for dynamic radio resource allocation and T_1 (=4) denotes short timescale for user scheduling, where $T = 2T_1$.

We use O to denote the set of services, U to denote the user set and $|U|$ to denote the total number of users. U is partitioned into two user subsets according to service types, i.e., u_R and u_M where $|U| = |u_R| + |u_M|$. u_R and u_M represent the uRLLC and eMBB service devices existing in the network respectively and $|u_R|$ and $|u_M|$ represent the number of uRLLC and eMBB devices respectively. The total number of bandwidth resources is represented by BW. $BW_R(t_k)$ and $BW_M(t_k)$ are two resource subsets allocated to uRLLC and eMBB services at the time slot t_k, respectively, where $BW = BW_R(t_k)+BW_M(t_k)$. We assume the allocated bandwidth will not change during the entire time period T.

Moreover, let C represent the set of sub-channels and $|C|$ represent the number of sub-channels, which is defined as the smallest resource unit allocated to transmission traffic over a period of time. For simplicity, we assume that the channel conditions are flat over entire BW during the entire time period T and each sub-channel has the same bandwidth, expressed as $\frac{BW}{|C|}$. Therefore, the bandwidth $BW_R(t_k)$ and $BW_M(t_k)$ are determined by the number of sub-channels. We use $C_o(t_k)$ to denote the number of sub-channels allocated to the service $o \in O$ at the time slot t, then $C_R(t_k)$ and $C_M(t_k)$ represent the number of sub-channels allocated to uRLLC and eMBB services respectively. By the way, o denotes one of the service of uRLLC or eMBB. In this regard, the bandwidth of each service o can be represented as $BW_o(t_k) = \frac{BW}{|C|} C_o(t_k)$.

In the data transmission process, both of the uRLLC services and eMBB services are considered as the FDMA (Frequency Division Multiple Access)-like strategies. At beginning time slot t_k of $k - th$ long timescale, all users are equally allocated the total bandwidth. Thus, the allocated bandwidth of any user u requesting uRLLC or eMBB service o is $BW_{o,u}(t_k)(= \frac{BW_o(t_k)}{|u_o|})$. In this equation, O represents the service set of uRLLC and eMBB and $o \in O$ represents a kind of services (uRLLC or eMBB services).

To better describe the sporadic uRLLC services, we define a binary variable $b_{R,u_R}(t)$ to illustrate the traffic status of uRLLC device u_R, i.e., $b_{R,u_R}(t) \in \{0, 1\}$. $b_{R,u_R}(t) = 1$ indicates that the IoT device u_R generate traffic for scheduling uRLLC services at time slot t, and $b_{R,u_R}(t) = 0$ otherwise. In particular, the expression of uRLLC transmission rates must additionally consider extreme latency and reliability requirements. We additionally consider the short packet structure and transmission error rate of the uRLLC traffic, which is different from Shannon's capacity formula [26] that eMBB services follow. The transmission rate for the uRLLC device u_R and eMBB device u_M is given by (1) and (2) respectively.

$$\begin{aligned} &r_{R,u}\big(BW_R(t_k), P_{R,u}(t), t\big) \\ &= b_{R,u}(t) \cdot BW_{R,u}(t_k)\left\{ \log_2\left[1 + \delta^c_{R,u}(t)\right] - \sqrt{\frac{V_k}{n}} Q^{-1}(\varepsilon^c_k)\right\} \\ &= b_{R,u}(t) \cdot \frac{C_R(t_k)}{|C|} \frac{BW}{|u_R|}\left\{ \log_2\left[1 + \delta^c_{R,u}(t)\right] - \psi\right\}, \qquad (1) \end{aligned}$$

$$\begin{aligned} &r_{M,u}\big(BW_M(t_k), t\big) \\ &= BW_{M,u}(t_k) \log_2\left\{1 + SN^c_{M,u}(t)\right\} \\ &= \frac{C_M(t_k)}{|C|} \frac{BW}{|u_M|} \log_2\left\{1 + SN^c_{M,u}(t)\right\}, \qquad (2) \end{aligned}$$

where $\delta^c_{R,u}(t) = \frac{g^c_{R,u}(t)P_{R,u}(t)}{\sigma^2}$, $\delta^c_{M,u}(t) = \frac{g^c_{M,u}(t)P_{M,u}(t)}{\sigma^2}$. $P_{R,u}(t)$ and $P_{M,u}(t)$ are the transmit power of uRLLC and eMBB services, where $P_{R,u}(t)$ can be controlled by the system for power consumption and $P_{M,u}(t)$ is fixed. $g^c_{R,u}(t)$ and $g^c_{M,u}(t)$ are channel gain-to-noise ratio of two kind of services. σ is the noise power. We use ψ to denote the reliability factor, which is expressed by $\psi = \sqrt{V_k/n}Q^{-1}(\varepsilon^c_k)$. We consider the short packet structure of the uRLLC traffic, so that the maximum transmission rate is related to the transmission error rate. $V_k = 1 - \frac{1}{[1+\delta^c_{R,u}(t)]^2}$ denotes the so-called channel dispersion which can be approximated to 1 [27]. It measures the random variation of the channel relative to the deterministic channel of the same capacity. $Q^{-1}(\cdot)$ denotes the inverse of the Gaussian Q function and ε^c_k denotes the transmission error rate. Due to the high reliability and ultra-low latency requirements, in order to avoid re-transmission, the transmission error rate ε^c_k should be set to a low threshold, thereby ψ can be considered as a constant. Moreover,we assume that rates for uRLLC and eMBB services are bounded as $r_{R,u}(BW_{R,u}(t_k), P_{R,u}(t), t) \leq r_{R,max}$ and $r_{M,u}(BW_{M,u}(t_k), t) \leq r_{M,max}$, respectively.

B. OVERALL NETWORK UTILITY

According to the service type, we assume that each kind of slice is assigned a private share, reflecting benefits that the slice obtains from a given resource allocation. In this setting, the design principle of resource allocation consists of two points. I) RAN resources are shared among services according to the private share. II) The resources are also fairly shared among devices of the service.

Let O denote the service set and S_o denote the private share for service $o \in O$, so that $\sum_{o\in O} S_o = 1$. Private shares are assigned for weighting according to communication requirements of frequency bandwidth and traffic flow, which guides the fair resource allocation and network utilization. A larger private share means a higher priority for radio resource. For instance, we assign a larger private share for video call services in order to give it a higher priority in occupying resources (e.g. bandwidth), which means the quality requirement of image or speech can be better ensured. We define the overall network utility $N(t)$ as the sum of the services' utilities $F_o(t)$ weighted by the private share. The network utility is defined as follows:

$$N(t) = \sum_{o\in O} S_o \cdot F_o(t). \qquad (3)$$

The service utility $F_o(t)$ is the sum of devices' utilities, resources of each service are fairly shared among devices. According to the fairness criteria [6], [7], [22], a service utility is logarithmic in its resource/bandwidth, expressed by:

$$\begin{aligned} F_o(t) &= \sum_{u\in U_o} \phi_u \cdot f_u\big(BW_{o,u}(t_k)\big) \\ &= \phi_u \sum_{u\in U_o} \log\left(BW_{o,u}(t_k)\right), \qquad (4) \end{aligned}$$

where U_o denotes the device set of service o and ϕ_u is the relative priority of device $u \in U_o$. For simplify, we define that ϕ_u is the same amongst all devices of the service o, such that $\phi_u \geq 0$ and $\sum_{u\in U_o} \phi_u = 1$. Therefore, $\phi_u = 1/|U_o|$. Combining the equations (3) and (4), we can rewrite the network utility as follows:

$$\begin{aligned} N(t) &= \sum_{o\in O} \sum_{u\in U_o} \frac{S_o}{|U_o|} \cdot \log\left(BW_{o,u}(t_k)\right) \\ &= \sum_{o\in O} \sum_{u\in U_o} \omega_{(o,u)} \cdot \log\left(BW_{o,u}(t_k)\right), \qquad (5) \end{aligned}$$

where $\omega_{(o,u)}$ denotes the device weight of device u, defined as the private share of service o which is divided by the current number of the IoT devices. That is, the device weight ω_o is equal amongst its current devices, i.e., $\omega_o = S_o/|U_o|$. We use N^{max} to denote the maximum network utility at the time slot t and the degree of network utility is then defined in a normalized form as:

$$H(t) = \frac{N(t)}{N^{max}} = \sum_{o\in O}\sum_{u\in U_o} \frac{\omega_o \cdot \log\left(BW_{o,u}(t_k)\right)}{N^{max}}. \tag{6}$$

C. LATENCY OF QUEUEING

For each time slot t, new random traffic is generated by each terminal and certain traffic can also be transmitted by the base station. Suppose that $A_u(t)$ denotes the amount of data generated by device u and $r_u(t)$ denotes the amount of data transmitted by device u, the queue backlog of device u is then denoted by $\mathbf{Q}_u(t)$ at the time slot t. Obviously, the updated queue backlog $\mathbf{Q}_u(t+1)$ at time slot $t+1$ can be derived by queue backlog $\mathbf{Q}_u(t)$, generated data $A_u(t)$ for processing and transmitted data $r_u(t)$ at the previous time slot t. If the amount of data that can be transmitted in a time slot is greater than the sum of the backlogged and generated data from the previous time slot, the current queue backlog will be none. On the contrary, when the transmission rate of the base station is insufficient, or amounts of connected devices and generated data are both large, the queue backlog will increase over time until the system crashes. In this setting, we use $\mathbf{Q}_{o,u}(t+1)$ to denote the queue backlog of device u for service o and $A_{o,u}(t)$ to denote the amount of data generated of device u for service o. Among that, o denotes one kind of uRLLC services or eMBB services (i.e., $A_{R,u}(t)$ and $A_{M,u}(t)$). The queues of two heterogeneous services both updates over time based on the following queuing dynamics:

$$\mathbf{Q}_{o,u}(t+1) = \max\left[\mathbf{Q}_{o,u}(t) - r_{o,u}(t), 0\right] + A_{o,u}(t) \quad (t \in 0, 1, 2, \ldots). \tag{7}$$

Extending from the standard Little's Law [28], the average transmission latency is proportional to the average queue length. Nevertheless, relying merely on the average queue length fails to account for the extreme value of queue length constraints. In [29] and [30], a probabilistic distribution of the queue length is proposed to solve the extreme latency and reliability constraint, however, the hard latency is still possible to be not satisfied. To better demonstrate and assess the extreme latency of uRLLC services, hard latency is proposed. With the maximum queue length, each calculation can constrain the worst-case latency. According to the actual communication situation, we assume that the uRLLC traffic is sporadic and follows a specific distribution with a threshold for traffic control. In other words, there is an upper limit to the traffic generated by each uRLLC device, and the absolute latency is expressed as:

$$T_{o,u}(t) := \frac{q_{o,u}^{max}(t)}{r_{o,u}(t)} \tag{8}$$

where $q_{o,u}^{max}(t)$ is the maximum queue length of device u for service o that appears at the time slot t. And $T_{o,u}(t)$ is the hard latency. Considering the ultra-low latency requirement for uRLLC services and high throughput requirements for eMBB services, the constraint of above queue backlog is:

$$\begin{aligned}\mathbf{Q}_{o,u}(t) \le q_{o,u}^{max}(t) &\le Q_o^{max} - A_{o,u}(t)\\ &= r_{o,u}(t)T_o^{max} - A_{o,u}(t),\end{aligned} \tag{9}$$

where Q_o^{max} represents the maximum length of queue backlog that service o can tolerate, and T_o^{max} represents the maximum latency threshold of the service o. We assume that the system can estimate the unfinished traffic data in their queues accurately [31]. Throughout the paper, if the queue $\mathbf{Q}_{o,u}(t)$ subjects to the definition of queue stability in (10), we can prove that the queue is strongly stable [32]. That is, if the maximum length of queue backlog time is bounded, the worst-case latency is limited as well.

$$\limsup_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left\{\mathbf{Q}_{o,u}(t)\right\} < \infty. \tag{10}$$

IV. PROBLEM STATEMENT AND ALGORITHM DESIGN

In this section, we formulate the optimization problem that will drive (i) the guarantee of ultra-low latency deadline and reduction of power consumption for sporadic uRLLC traffic and (ii) the improvement of service quality for eMBB services. Therefore, we formulate the problem model by minimizing the cost function and design the algorithm with Lyapunov optimization. Besides, we propose the optimal solution in terms of resource allocation decision strategy, optimal RAN resource allocation and service schedule.

A. PROBLEM FORMULATION

For the sporadic uRLLC service, RAN transmission must guarantee ultra-low latency (1ms) and high reliability while reducing the power consumption of IoT devices as much as possible. On the other hand, the eMBB service aims at maximizing the service quality for all users with high throughput. Since fulfilling the ultra-low latency requirements comes at the expense of a degraded QoS of eMBB services, we employ queue backlog to reasonably schedule the relationship among latency, power consumption, and network utility. Motivated by this, we formulate the cost function and propose the problem with the objective of minimizing the total cost. The optimization problem is then formulated as:

$$(\boldsymbol{P}): \min_{(BW_o, P_R, A_M)} \lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left\{c(t)\right\}, \tag{11}$$

$$\lim_{T\to\infty} \sup \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left\{\mathbf{Q}_{R,u}(t) + \mathbf{Q}_{M,u}(t)\right\} < \infty, \tag{12}$$

$$\begin{gathered}BW_R \ge BW_R^{min}, BW_M \ge BW_M^{min},\\ BW_R + BW_M = BW,\end{gathered} \tag{13}$$

$$q_{R,u}^{max}(t) \le \mathbf{Q}_R^{max} - A_{R,u}(t), q_{M,u}^{max}(t) \le \mathbf{Q}_M^{max} - A_{M,u}(t), \tag{14}$$

where $c(t)$ denote the cost function, formulated as the sum of average power consumption of uRLLC devices and the negative value of the network utility, i.e., $c(t) = \sum_{u \in U_R} \frac{P_{R,u}(t)}{P_R^{max}} b_{R,u}(t) - H(t)$. Furthermore, $BW_o = \big(BW_R(t), BW_M(t)\big)_{t=0}^{\infty}$, $P_R = \big(P_{R,u}(t), \forall u \in U_R\big)_{t=0}^{\infty}$, $A_M = \big(A_{M,u}(t), \forall u \in U_R\big)_{t=0}^{\infty}$ and BW_R^{min} and BW_M^{min} denote the minimum bandwidth requirements for uRLLC and eMBB services, respectively. Constraint (12) is queue stability condition, constraint (13) guides the bandwidth allocation, and constraint (14) restricts the latency from exceeding the maximum latency threshold. Where $q_{R,u}^{max}(t)$ and $q_{M,u}^{max}(t)$ represent the maximum queue length of uRLLC and eMBB devices u at the time slot t, respectively. $\mathbf{Q}_R^{max}$ and $\mathbf{Q}_M^{max}$ represent the maximum queue backlogs that uRLLC and eMBB services can tolerate, respectively. In this problem, we need to determine the bandwidth allocation, transmit power and service quality for each IoT device at each time slot. However, in general, $\boldsymbol{P}$ is difficult to solve as it is a stochastic optimization problem and optimal decisions are temporally correlated [19].

B. ALGORITHM DESIGN

Since the queue state is coupling among different time slots, achieving the above goal will be difficult by the conventional convex optimization [33]. Thus, we employ Lyapunov optimization to derive the bandwidth allocation, latency guarantee, transmit power control and service quality decision algorithms. It takes two timescales to obtain analytical performance boundaries inherited from [14]. In this regard, we first define the quadratic Lyapunov function $L(t)$ as:

$$L(t) \triangleq \frac{1}{2}\Big\{ \sum_{u \in U_R} \mathbf{Q}_{R,u}^2(t) + \sum_{u \in U_M} \mathbf{Q}_{M,u}^2(t) \Big\}. \tag{15}$$

Let $\mathbf{Q}(t_k) = \{\mathbf{Q}_{R,u}(t), u \in U_R; \mathbf{Q}_{M,u}(t), u \in U_M\}, \forall t \in [t_k, t_k + T - 1]$ be the queue backlog vector. Thus, the Lyapunov drift function which is defined as the expected change over the time period T can be written as:

$$\Delta_T\big(L(t_k)\big) \triangleq \mathbb{E}\Big\{ L(t_k + T) - L(t_k) | \mathbf{Q}(t_k) \Big\}. \tag{16}$$

Next, we use the penalty function which is the sum of cost functions to design Lyapunov drift-plus-penalty function for uRLLC and eMBB services at time slot t. The Lyapunov drift-plus-penalty function can be expressed as:

$$\tau(t_k) = \Delta_T\big(L(t_k)\big) + V\,\mathbb{E}\Big\{ \sum_{t=t_k}^{T_k+T-1} c(t) | \mathbf{Q}(t_k) \Big\}, \tag{17}$$

where V is a control parameter in the proposed algorithm, used to balance the cost function (i.e., power consumption and network utility) and hard latency. For instance, when the control parameter V is large, the system will prefer to reduce the transmit power and improve service quality rather than optimize the latency. We then derive an upper-bound of $\tau(t_k)$ using queueing dynamics under any feasible $BW_o(t)$, $P_{R,u}(t)$ and $A_{M,u}(t)$, as specified in the following lemma.

Lemma 1: Let $V > 1$ and $t_k = kT$ where $k = 0, 1, \ldots$ For arbitrary $BW_o(t)$, $P_{R,u}(t) \in [0, P_{R,max}]$ and $A_{M,u}(t) \in [0, A_{M,max}]$, we have:

$$\begin{aligned}
&\tau(t_k) \\
&\leq cons_1 \cdot T + V\,\mathbb{E}\Big\{ \sum_{t=t_k}^{T_k+T-1} c(t) | \mathbf{Q}(t_k) \Big\} \\
&\quad - \mathbb{E}\Big\{ \sum_{t=t_k}^{T_k+T-1} \sum_{u \in U_R} \mathbf{Q}_{R,u}(t) \Big[r_{R,u}\big(BW_R(t_k), P_{R,u}(t), t\big), t\big) \\
&\quad - A_{R,u}(t) \Big] | \mathbf{Q}(t_k) \Big\} - \mathbb{E}\Big\{ \sum_{t=t_k}^{T_k+T-1} \sum_{u \in U_M} \mathbf{Q}_{M,u}(t) \\
&\quad \Big[r_{M,u}\big(BW_M(t_k), t\big) - A_{M,u}(t) \Big] | \mathbf{Q}(t_k) \Big\},
\end{aligned} \tag{18}$$

where $cons_1 \triangleq \big(|U_R|(A_{R,max}^2 + r_{R,max}^2) + |U_M|(A_{M,max}^2 + r_{M,max}^2)\big)/2$, and $A_{R,max}$ is the fixed maximum generated traffic of uRLLC services.

Proof: The proof is similarly based on the principle of queuing dynamics in (7), which is omitted due to space limitation. ■

One of the design intents of Lyapunov optimization is to determine a control action for minimizing the right hand side (R.H.S.) of (18). However, the prior knowledge of future queue backlog may not always be available, which depends on the task generated processes $A_{o,u}(t)$, resource allocation solutions, and time-varying network conditions. In this regard, we take an approximation to assume the future values of queue backlogs $\mathbf{Q}_{R,u}(t)$ and $\mathbf{Q}_{M,u}(t)$ during $[t_k, t_k + T - 1]$ are roughly equal current values, i.e., $\mathbf{Q}_{R,u}(t) \approx \mathbf{Q}_{R,u}(t_k)$ and $\mathbf{Q}_{M,u}(t) \approx \mathbf{Q}_{M,u}(t_k)$ for all $t_k \leq t \leq t_k + T - 1$ [14]. The upper bound of drift-plus-penalty $\tau(t_k)$ "loosen" as shown in the following lemma.

Lemma 2: Let $t_k = kT$ where $k = 0, 1, \ldots$ Under any feasible value of $BW_o(t)$, $P_{R,u}(t)$ and $A_{M,u}(t)$, we have:

$$\begin{aligned}
&\tau(t_k) \\
&\leq cons_2 \cdot T + V\,\mathbb{E}\Big\{ \sum_{t=t_k}^{T_k+T-1} c(t) | \mathbf{Q}(t_k) \Big\} \\
&\quad - \mathbb{E}\Big\{ \sum_{t=t_k}^{T_k+T-1} \sum_{u \in U_R} \mathbf{Q}_{R,u}(t) \Big[r_{R,u}\big(BW_R(t_k), P_{R,u}(t), t\big), t\big) \\
&\quad - A_{R,u}(t) \Big] | \mathbf{Q}(t_k) \Big\} - \mathbb{E}\Big\{ \sum_{t=t_k}^{T_k+T-1} \sum_{u \in U_M} \mathbf{Q}_{M,u}(t) \\
&\quad \Big[r_{M,u}\big(BW_M(t_k), t\big) - A_{M,u}(t) \Big] | \mathbf{Q}(t_k) \Big\},
\end{aligned} \tag{19}$$

where $cons_2 \cdot T \triangleq cons_1 + (T-1)\big(|U_R|(A_{R,max}^2 + r_{R,max}^2) + |U_M|(A_{M,max}^2 + r_{M,max}^2)\big)/2$.

Proof: The proof is similarly based on the principle of queuing dynamics in (7), which is omitted due to space limitation. ■

With minimizing R.H.S. of (19) in terms of $BW_o(t)$, $P_{R,u}(t)$ and $A_{M,u}(t)$, we can determine the bandwidth allocation for two services, which are latency and power consumption control for uRLLC and service quality decision for eMBB. When determining bandwidth allocation of the $k-th$ long timescale at time slot t_k, the future channel gain states during $[t_k, t_k+T-1]$ are not able to know. Due to the wireless channel quality is rarely changed over short time period [13], statistics of channel states are similar with the most recent ones. Therefore, we address the above issue by approximating the future channel gain-to-noise radio $g^c_{R,u}$ and $g^c_{M,u}$ during the new time period $[t_k, t_k+T-1]$ are the same as the current value [12].

C. PERFORMANCE ANALYSIS

In this subsection, we will provide a theoretical result, which analyzes optimal bounds on the power and service quality performance (average cost) of our proposed problem over all stable queue. In such theoretical algorithm, we assume that future information such as the channel gains is known in advance and achieve the optimum by exhaustive searching, called semi-offline algorithm. The performance is characterized in Theorem 1.

Theorem 1: Suppose there exists an $\epsilon > 0$ such that $\overline{A} + \epsilon 1 \in \Lambda$, where $\overline{A} = (A_1, A_2, \ldots, A_u)$ represents a given job generated rate vector and has $A_u = \mathbb{E}\{A_u(t)\}$. $A_u(t)$ is the generated rate of device u at time slot t. Λ denotes the capacity region of the system – i.e., the closure of set of rates $\overline{A}$ to ensure the queue stability expressed in (10). 1 is the vector of all 1's. Then, we have:

$$\overline{\mathbf{Q}}_T \triangleq \limsup_{K\to\infty} \frac{1}{K}\sum_{k=0}^{K-1}\left\{\sum_{u=1}^{|U_R|}\mathbb{E}\{\mathbf{Q}_{R,u}(kT)\} + \sum_{u=1}^{|U_M|}\mathbb{E}\{\mathbf{Q}_{M,u}(kT)\}\right\} \le \frac{cons_2 + Vc_{max}}{\epsilon},$$

$$c_{semi} \triangleq \limsup_{K\to\infty} \frac{1}{K}\mathbb{E}\{c(t)\} \le c^* + \frac{cons_2}{V}, \tag{20}$$

where k denotes the number of time period T, which achieves $t_k = kT$ for all $k = 0, 1, \ldots, K-1$. c_{max} denotes the maximum cost for the traffic arrival rates $\overline{A}$, c^* denotes the minimum time-average cost for all users, and c_{semi} denotes the optimal cost of semi-offline scheme.

Proof: See [34], which is omitted due to space limitation. ■

D. OPTIMAL SOLUTION

The resource allocation decision for heterogeneous services is made through solving the optimization problem to minimize R.H.S. of (19). The exhaustive search scheme is an approach, however, it needs to solve all possible combinations with high computational complexity. In this regard, we design a two timescale framework and propose a more computationally efficient sub-optimal scheme to address the allocation decision by an efficient alternative minimization algorithm. The two-timescale algorithm makes a decision on bandwidth allocation every T time slot. Service controlling and scheduling (i.e., latency and power consumption control for uRLLC services and quality decisions for eMBB services) are made at each short time slot. Thus, we first design the resource allocation decision strategy according to the stabilized threshold. Then we introduce the bandwidth allocation algorithm and two sub-algorithms for service control.

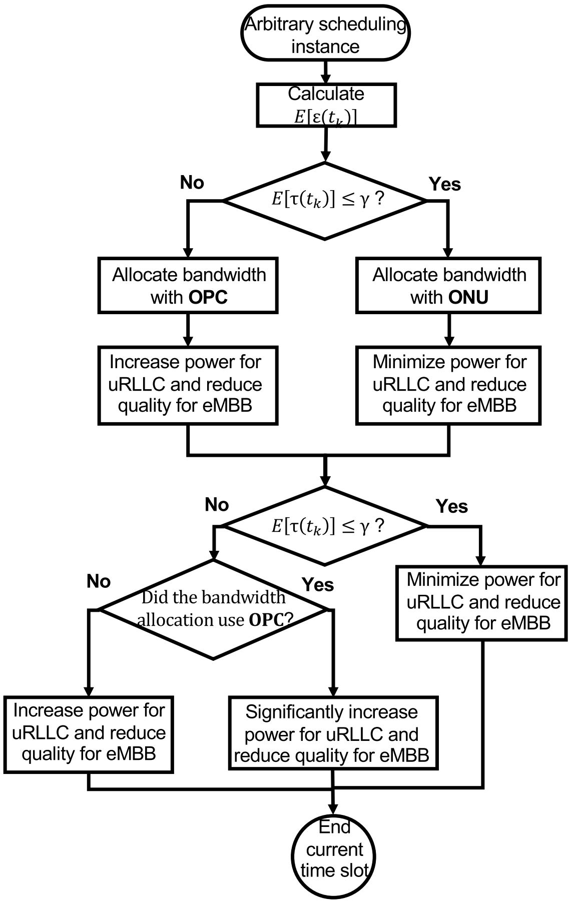

Arbitrary scheduling instance
Calculate $E[\varepsilon(t_k)]$
No
$E[\tau(t_k)] \le \gamma$?
Yes
Allocate bandwidth with **OPC**
Allocate bandwidth with **ONU**
Increase power for uRLLC and reduce quality for eMBB
Minimize power for uRLLC and reduce quality for eMBB
No
$E[\tau(t_k)] \le \gamma$?
Yes
No
Did the bandwidth allocation use **OPC**?
Yes
Minimize power for uRLLC and reduce quality for eMBB
Increase power for uRLLC and reduce quality for eMBB
Significantly increase power for uRLLC and reduce quality for eMBB
End current time slot

FIGURE 3. Flow diagram of proposed alternative minimization algorithm.

1) RESOURCE ALLOCATION DECISION STRATEGY

According to the Lyapunov stability theorem and Lyapunov's second method, the system can reach a stable state when $\triangle_T(L(t_k))$ can be stabilized within a certain range. Assume that the stability threshold for ensuring the normal operation of system is γ, we can visually determine whether the system has remaining network resources to further optimal latency or reduce power consumption and improve service quality. The relationship is expressed as follows:

$$\mathbb{E}\left[\tau(t_k)\right] \le \gamma. \tag{21}$$

Fig. 3 shows our alternative minimization algorithm based on Lyapunov optimization. For every T time slot (long timescale), if the drift-plus-penalty function satisfies the condition of (21), it indicates that the system is stable and may have remaining resources based on the current allocation decision. At this time, the network determines the optimal network utility (ONU) scheme as the bandwidth allocation policy, and can appropriately reduce the power consumption for uRLLC devices and improve the service quality for eMBB users. On the contrary, if the value of drift-plus-penalty function is greater than the stability threshold (i.e., $\mathbb{E}\left[\tau(t_k)\right] > \gamma$), the queue backlog is large and the system is unstable, which may cause system failure. The system then uses the optimal power consumption (OPC) scheme to allocate bandwidth and appropriately increases the power consumption and reduces the service quality to stabilize the queue backlog. For each short timescale, notice that determining the control and scheduling policy needs to both compare with the threshold γ and consider the bandwidth allocation policy in the current period. When $\mathbb{E}\left[\tau(t_k)\right] > \gamma$ and the bandwidth allocation policy is OPC, the system should significantly increase power for uRLLC devices and reduce the service quality for eMBB users, enabling the system to stabilize as fast as possible.

2) OPTIMAL BANDWIDTH ALLOCATION

For the bandwidth allocation, the proposed algorithm is summarized in Algorithm 1 to solve the deterministic optimization problem at each T time slot. We first assign the same bandwidth at the beginning time slot as there is no previous network information for reference. It is not difficult to identify that the optimization problem is monotonous in terms of bandwidth, and then the optimal solution is $(BW_R^{min}, BW - BW_R^{min})$ or $(BW - BW_M^{min}, BW_M^{min})$. In order to directly search for the optimal solution, we design two resource allocation policies to find the solution $\big(BW_R(t_k), BW_M(t_k)\big)$ by an alternative manner. The allocation policies are OPC scheme and ONU scheme, which are determined depending on the stability threshold γ. If the system determines the OPC scheme, we calculate the corresponding temporary power consumption $P_R^{opt,1}$ and $P_R^{opt,2}$ by LTPC algorithm in the case of $BW_R = BW_R^{min}$ or $BW_R = BW - BW_M^{min}$, respectively. After calculating $D^{opt,1}$ and $D^{opt,2}$ at two possible combination of bandwidth allocation $(BW_R^{min}, BW - BW_R^{min})$ and $(BW - BW_M^{min}, BW_M^{min})$ in step 7 of Algorithm 1, we make a decision for bandwidth allocation, and the formulations of $D^{opt,1}$ and $D^{opt,2}$ are expressed by:

$$\begin{aligned} D^{opt,1} = & \sum_{u\in U_R} BW_{R,u}^{min} \mathbf{Q}_{R,u^{opt,1}(t)}(t_k)\Big(\log_2 \big(1 \\ & + SN^c_{R,u^{opt,1}}(t)\big) - \psi\Big) \\ & + \sum_{u\in U_M} BW_{M,u} \mathbf{Q}_{M,u}(t_k) \log_2 \left(1 + SN^c_{M,u}(t)\right), \end{aligned} \tag{22}$$

$$\begin{aligned} D^{opt,2} = & \sum_{u\in U_R} BW_{R,u} \mathbf{Q}_{R,C(t)}(t_k)\Big(\log_2 \big(1 \\ & + SN^c_{R,u^{opt,2}}(t)\big) - \psi\Big) \\ & + \sum_{u\in U_M} BW_{M,u}^{min} \mathbf{Q}_{M,u}(t_k) \log_2 \left(1 + SN^c_{M,u}(t)\right), \end{aligned} \tag{23}$$

where $u^{opt,1}$ and $u^{opt,2}$ denote the device which are scheduled with the corresponding transmit power $P_{R,u^{opt,1}}(t)$ and $P_{R,u^{opt,2}}(t)$ respectively. On the other hand, if the system determines the ONU scheme, we calculate the $W^{opt,1}$ and $W^{opt,2}$ as two possible combination of bandwidth allocation $(BW_R^{min}, BW - BW_R^{min})$ and $(BW - BW_M^{min}, BW_M^{min})$ in step 10,11 of Algorithm 1 to make a decision for bandwidth allocation.

Algorithm 1 Bandwidth Allocation Algorithm.

```
Input  : Queue backlog $\mathbf{Q}_{o,u}(t_k)$, stability threshold $\gamma$.
Output: Optimal solution $BW_o(t_k)$.
1  for every time slot $t_k = kT, k = 0, 1, 2, \ldots$ do
2      if $k = 0$ then
3          $BW_R(t_k) = BW/2$ and $BW_M(t_k) = BW/2$.
4      else
5          Assume the channel gain-to-noise radio $g^c_{R,u}$ and
           $g^c_{M,u}$ during the new time period $[t_k, t_k + T)$ are
           the same as that during $[t_k - T, t_k)$.
6      end
7      if $\left[\tau(t_k)\right] > \gamma$ then
8          Apply Algorithm 2 to calculate the bandwidth
           allocation solution $\big(BW_R(t_k), BW_M(t_k)\big)$.
9      end
10     else if $\left[\tau(t_k)\right] \leq \gamma$ then
11         Apply Algorithm 3 to calculate the bandwidth
           allocation solution $\big(BW_R(t_k), BW_M(t_k)\big)$.
12     end
13 end
```

Algorithm 2 Optimal Power Consumption (OPC) Algorithm.

```
Input  : $\mathbf{Q}_{o,u}(t_k)$, $BW_R^{min}$, $BW_M^{min}$
Output: Optimal solution $BW_i$.
1  for every time slot $t_k = kT, k = 0, 1, 2, \ldots$ do
2      Apply LTPC to calculate the temporary power
       consumption $P_R^{opt,1}$ and $P_R^{opt,2}$ when $BW_R = BW_R^{min}$
       and $BW_R = BW - BW_M^{min}$, respectively.
3      Calculate $D^{opt,1}$ and $D^{opt,2}$ by formulation (22) and
       (23) respectively,
4      if $D^{opt,1} \geq D^{opt,2}$ then
5          $BW_R = BW_R^{min}, BW_M = BW - BW_R^{min}$.
6      else
7          $BW_R = BW - BW_M^{min}, BW_M = BW_M^{min}$.
8      end
9  end
```

3) OPTIMAL LATENCY, TRANSMIT POWER AND SERVICE QUALITY

For each bandwidth allocation case, we obtain the short timescale solutions for the service controlling and scheduling.

Algorithm 3 Optimal Network Utility (ONU) Algorithm.

```
Input  : $BW_R^{min}$, $BW_M^{min}$, $\omega_{i,u}$
Output: Optimal solution $BW_i$.
for every time slot $t_k = kT, k = 0, 1, 2, \dots$ do
    Apply SQD to calculate $A_{M,u}^{opt,1}$ and $A_{M,u}^{opt,2}$ when $BW_R = BW_R^{min}$ and $BW_R = BW - BW_M^{min}$, respectively.
    Calculate $W^{opt,1} = \sum_{u \in U_R} \omega_{R,u} \log(BW_R^{min}) + \sum_{u \in U_M} \omega_{M,u} \log(BW - BW_R^{min})$, and $W^{opt,2} = \sum_{u \in U_R} \omega_{R,u} \log(BW - BW_M^{min}) + \sum_{u \in U_M} \omega_{M,u} \log(BW_M^{min})$.
    if $W^{opt,1} \geq W^{opt,2}$ then
        $BW_R = BW_R^{min}, BW_M = BW - BW_R^{min}$.
    else
        $BW_R = BW - BW_M^{min}, BW_M = BW_M^{min}$.
    end
end
```

By dividing the problem for minimizing R.H.S. of (19) into two sub-algorithms, the latency and transmit power control (LTPC) for uRLLC service and service quality decision (SQD) for eMBB services, the bandwidth allocation problem can be solved. Since the problem is jointly convex with respect to $P_{R,u}(t)$ and $A_{M,u}(t)$ and the feasible region is the Cartesian product of those variables. Our algorithm can be guaranteed to converge to the global optimal solution, termed as the Gauss-Seidel method in literature [23].

a: LATENCY AND TRANSMIT POWER CONTROL (LTPC)

The LTPC algorithm controls transmit power to schedule hard latency and power consumption. After decoupling SP_1 from the R.H.S. of (19), the optimal transmit power for uRLLC devices can be obtained with a fixed bandwidth allocation solution $\big(BW_R(t_k), BW_M(t_k)\big)$ by solving:

$$(\boldsymbol{SP_1}) : \min V \frac{P_{R,u}(t)}{P_{R,max}} - \mathbf{Q}_{R,u}(t) r_{R,u}\big(P_{R,u}(t), t\big), \quad 0 \leq P_{R,u}(t) \leq P_{R,max}. \tag{24}$$

Thus, for every time slot $t \in [t_k, t_k + t - 1]$, the optimal solution of transmit powers is achieved at either the stationary point of the objective function SP_1 or one of the boundary points $(0, P_{R,max})$. For all uRLLC devices, optimal transmit powers are given in closed form by:

$$P_{R,u}^*(t) = \left\{ \max \left\{ \frac{P_R^{max} \mathbf{Q}_{r,u}(t) BW_R(t_k)}{\ln 2 \cdot V} - \frac{\sigma}{g_{R,u}^c(t)}, 0 \right\}, P_{R,max} \right\}, u \in U_R. \tag{25}$$

b: SERVICE QUALITY DECISION (SQD)

The SQD algorithm controls the service quality of eMBB services to schedule hard latency and service quality (e.g., the served content size). For every time slot $t \in [t_k, t_k + t - 1]$, we need to decide the service quality for each eMBB devices to maximize the follow metric:

$$(\boldsymbol{SP_2}) : \max \mathbf{Q}_{M,u}(t)\big(r_{M,u}(t) - A_{M,u}(t)\big), \quad 0 \leq A_{M,u}(t) \leq A_{M,max}. \tag{26}$$

With the fixed bandwidth allocation solution, SP_2 is a linear program in terms of $A_{M,u}(t)$, the optimal solution should be one of the boundary points similar to SP_1.

V. PERFORMANCE EVALUATION

A. SIMULATION SETUP

First, this section defines the main indicators for performance evaluation and introduces the related system simulation parameters, including the hard latency, the power consumption of IoT devices, and user satisfaction to evaluate service quality. The latency and power consumption have been defined in the previous section, so we only define performance indicators for quality of service, also known as user satisfaction. The degree of user satisfaction analyzes the value of $A_{M,u}(t)$ and is modeled as service quality $US_i(A_{M,u}(t))$ divided by user's personalized quality requirement US_i^{max} [5], expressed as follows:

$$M_i(A_{M,u}(t)) = \frac{US_i(A_{M,u}(t))}{US_i^{max}}, \tag{27}$$

where $US_i(A_{M,u}(t)) = \beta_i \log_2((A_{M,u}(t))$ denotes the service quality of eMBB user i and β_i reflects the characteristics of requested services, which follow peak signal-to-noise ratio [35].

In the simulation, we assume that the RAN scenario only contains one BS which provides radio resources for slices to support heterogeneous services. The total bandwidth is set to 100MHz. There are 3 kinds of uRLLC slices providing radio resource for 15 uRLLC devices and 3 kinds of eMBB slices for 15 eMBB users, which are located at a random distance between 10m and 500m from the BS. As for the channel gains, we use the small-scale fading channel model [23] which is exponentially distributed with unit mean.

For uRLLC services, the minimum bandwidth is set to 25MHz and the maximum transmit power is set to 1W [36]. The reliability factor ψ can be seen as a constant in formulating the transmission rate and set to 10^{-5}. According to 3GPP standard, the upper bound of RAN side latency threshold is set to 1ms. Each service device generates data randomly. The average arrival rate is set to 600kbps following Poisson distribution and the maximum average arrival rate is 1Mbps. For eMBB services, the minimum bandwidth is set to 50MHz and the typical value of transmit power is set to 8W [37]. The upper bound of RAN side latency threshold is set to 5ms. The finite set of service content size requested by the user is selected randomly between 3.5Mbits to 9Mbits, which is set according to the YouTube video streaming using popular H.264 codec. What's more, the private share is set based on the fraction of arrival rates of all devices at each time slot [22]. In addition, we first fix the timescale T for bandwidth allocation to be 240 time slots and run experiments

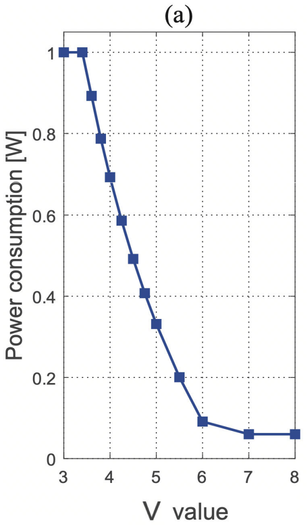


FIGURE 4. Avg. power cost for uRLLC services (a) /average degree of user satisfaction for eMBB services (b) vs. the control parameter *V*.

with different V values. Then we fix the trade-off parameter V to be $4\text{Mbit}^2 * \text{W}^{-1}$ and vary T from 30 time slots to 1080 time slots, which is a sufficient and efficient range for exploring the proposed algorithm.

B. NUMERICAL RESULTS

From the problem model, we can notice that the performance of proposed algorithm depends on control parameter V and timeslot parameter T. For all comparison schemes, we show average values (power consumption, user satisfaction and latency) over arrival data sets.

First, we analyze the relationship between the power consumption for uRLLC services (Fig. 4(a)) and/or the degree of user satisfaction for eMBB services (Fig. 4(b)) under different control parameter V with the fixed T $(= 240$ time slots) in Fig. 4. From Fig. 4(a), we can notice that as the parameter V goes from 3 to 8, the power consumption for uRLLC services reduces from 1W to about 0.09W. That is to say, if the control parameter V is sufficiently large, the power consumption will converge to the optimal power consumption and it reduces inversely proportional to the parameter V. When the parameter is set to the value smaller than 3, the power consumption is an average of around 1 W, which is the maximum power cost set to uRLLC services. The reason is that the proposed algorithm adjusts V to balance power consumption and latency reduction. A larger V means that the algorithm is suitable for a power-sensitive situation. Conversely, a smaller V means that it has stricter requirements on latency reduction. Thus, the algorithm schedules uRLLC services with the maximum power to achieve ultra-low latency and system stability. Meanwhile, as shown in Fig. 4(b), we observe that when V goes to infinity, the average degree of user satisfaction for eMBB services decreases almost linearly and it is regarded as unbounded until to minimum value. Because (i) our algorithm considers both satisfying the requirements of uRLLC and eMBB services, especially the ultra-low latency constraints, and (ii) our algorithm is not service quality conserving and can adjust the arrival rate (e.g. the video quality) according to the total queue backlog and power cost. In addition, we observed that the degree of user satisfaction peaks when V is small (< 3), since the algorithm schedules enough power for uRLLC services, leaving more bandwidth resources for eMBB services. Thus, the results of Fig. 4 verify the trade-off between the ultra-low latency and power consumption/service quality as shown in Theorem 1.

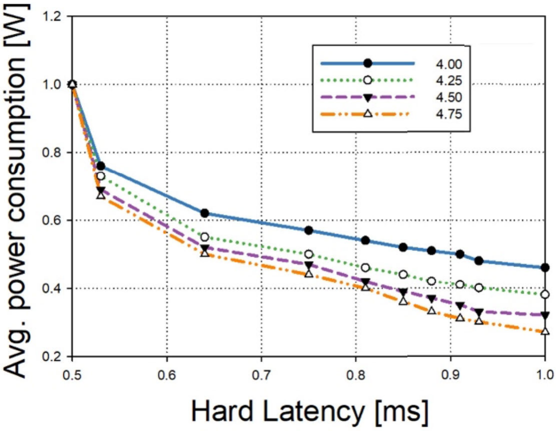


FIGURE 5. Power consumption vs. latency for uRLLC services.

In order to further evaluate power consumption and hard latency, we analyze the relationship between the power consumption and latency among different V value setting for uRLLC services. As shown in Fig. 5. We ignore the relationship for eMBB services, as latency constraints merely make a little sense. We can notice that the hard latency is inversely proportional to power consumption, which means that the latency increases as the power consumption decreases and the growth increase with V value. Thus, the proper V should be set at the beginning according to the actual scenario requirements. For instance, when the maximum latency threshold is set to about 15ms, $V = 4.25/4.5$ can be a good choice, since the power is low enough (far lower than the ones at $V = 4.00$ and very close to $V = 4.75$). On the other hand, the service quality also basically meets the requirements of eMBB users at this moment, whose average degree of user satisfaction exceeds 90%, as shown in Fig. 4(b).

When analyzing the impact of T value, we main consider three algorithms for comparison, BSRA (Bandwidth Slicing and Resource Allocation) algorithm [12], ACS (Alternative Concave Search) algorithm [13], and Loading Balancing algorithm [14]. BSRA also uses Lyapunov optimization to deal with the timescale problem, however, it pays more attention to power reduction for IoT devices and ignores hard latency. ACS transforms the optimization problem into a biconcave maximization problem, which considers the optimal bandwidth slicing and spectrum efficiency. In the Load Balancing algorithm, the amount of workload is proportional

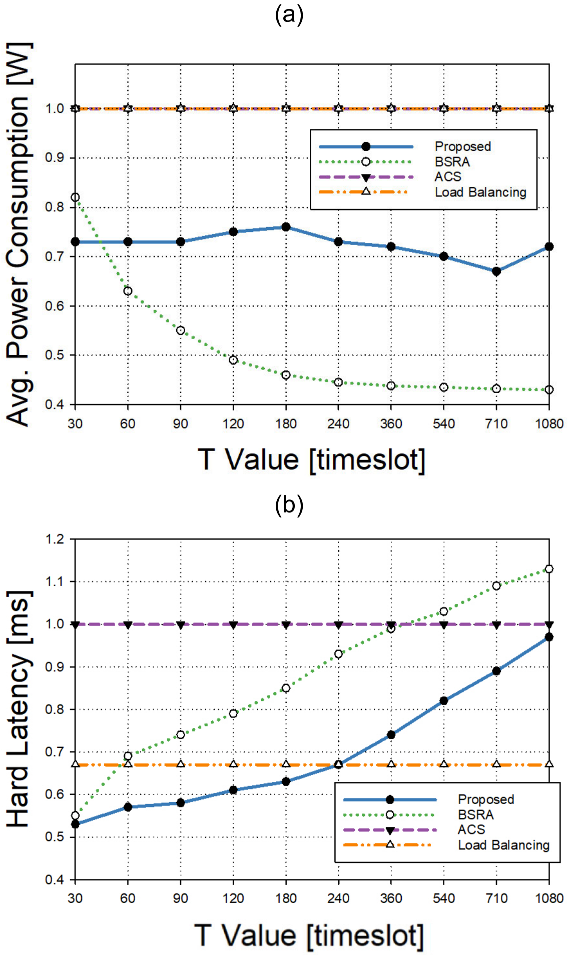


FIGURE 6. Avg. power cost/latency of all schemes under different *T* values.

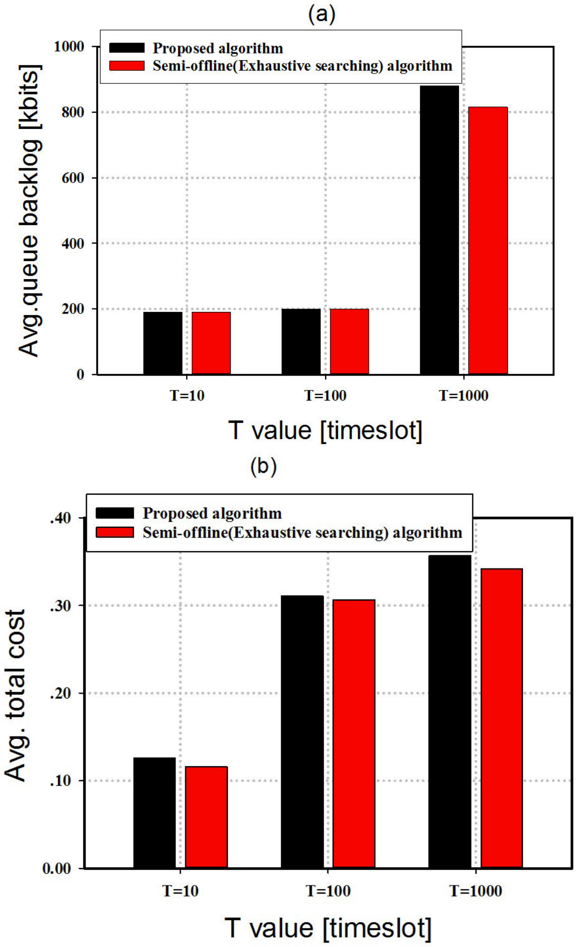


FIGURE 7. Avg. queue backlog/total cost for the proposed algorithm vs. semi-offline (exhaustive searching) algorithm in different long timescales.

to service capacity, regardless of power cost, so it should have good latency characteristics.

First, we vary T from 30 timeslots to 1080 timeslots and fix V to be 4.00. Corresponding results of different algorithms are shown in Fig. 6. In Fig. 6(a), we can notice that changing T has a relatively small effect on our proposed algorithm in terms of reducing power consumption, far less than that on ACS and Loading Balancing. The average fluctuation of our algorithm's power consumption is between 4.7% and 9.3%. On the contrary, the power consumption of BSRA varies significantly with T, which results in higher power reductions than the proposed algorithm. The phenomenon can be attributed to the scheduling principle of our algorithm: the key point is to ensure both the ultra-low latency requirement and the service quality while reducing the power consumption for IoT devices, rather than minimizing power consumption on a large scale. From Fig. 6(b), we note that T plays an important role in hard latency variation and may result in performance deterioration. Our algorithm has a lower latency than other schemes in most cases. In the extreme case, the hard latency increases proportionally and becomes high when $T = 1080$ time slots. This is not surprising - recall that the bound of queue size given in Theorem 1, the $cons2$ term is proportional to T, i.e., the latency increases with T. Similar to V, a proper T should also be chosen to balance the two desirable objectives according to actual scenario requirements. With a reasonable choice of T, for example, when T is 240, we can ensure that the power consumption is within an acceptable range while ensuring low latency. In Fig. 6(b), the BSRA algorithm does not have latency limitation for 1ms. When T is larger enough, the hard latency may exceed 1ms. Thus, compared with the BSRA algorithm, our proposed algorithm can better reflect the advantages in hard latency control.

Our proposed algorithm only uses current queue backlog and channel states, rather than future wireless information. The objective is to minimize the upper bound of cost with choosing the bandwidth allocation every T time slots, and choosing power consumption and service quality every single time slot. We assume there is a theoretical algorithm (called Semi-offline algorithm), which can obtain partial future wireless channel states and achieve the optimum over all stable queue by exhaustive searching, though it is not practically implementable. In such exhaustive searching algorithm, the optimal cost is the theoretical result analyzed in Theorem 1 and can be the optimal baseline for our proposed algorithm to obtain the performance gap. Fig. 7 compares our algorithm with the semi-offline (exhaustive searching) algorithm, in terms of queue backlog and system cost. Queue backlog is a key metric for hard latency and system stability, defined in the third chapter. The average system cost is the objective of our problem model, the sum of power consumption and negative network utility. With the same latency constraint, it can be seen that the proposed algorithm achieves similar performance to the theoretical result, especially when T is close to 100. Fig. 7(a) proves that our algorithm can effectively ensure the stability of the system queue and hard latency requirement, as the queue backlog is similar to the semi-offline (exhaustive searching) algorithm. When reaching $T = 1000$, the queue backlog length of our algorithm is longer and the cost is higher, from Fig. 7(b). According to our optimization goals, higher cost means lower performance. Thus, we can conclude that when it is greater than 1000T, the performance of our algorithm is lower than the optimal solution (baseline).

The above experience reveals that the proposed algorithm can perform better for uRLLC and eMBB hybrid services. Therefore, in this section, we use average power consumption, user satisfaction, hard latency and total cost to evaluate the performance. As the performance depends on parameters V and T, we first show corresponding results under different V values. It proved that our proposed algorithm can balance the relationship between power consumption and user satisfaction, i.e., when V is large, the algorithm outperforms scheme in power cost reduction. We then analyze the relationship between power consumption and latency for uRLLC services, which proves the hard latency guarantee. As for parameter T, it also plays an important role in the trade-off between latency and power cost. Finally, we compare our proposed algorithm with the theoretical one, which proves that our model achieves similar performance of exhaustive searching algorithm (the optimal solution).

VI. CONCLUSION

In this paper, we propose a dynamic resource allocation algorithm with RAN slicing and heterogeneous services scheduling, in order to ensure the extreme QoS of differentiated IoT services. We first present the mathematical model with the objective of minimizing total cost, formulated with the sum of power consumption of uRLLC devices and the negative value of the network utility. The problem model also obeys latency constraints and considers the error rate, which is used to guarantee the latency and reliability for uRLLC services. Then we employ the Lyapunov optimization to design a two timescale algorithm, long-timescale bandwidth allocation and short-timescale service control. Performance analysis was conducted for our algorithm, which explicitly characterizes the relationship between the control parameters and services performance, including power consumption and user satisfaction. With comparing, we proved that our algorithm outperforms BSRA, ACS and Loading Balancing algorithms in terms of hard latency and total cost.

REFERENCES

[1] A. Anand, G. De Veciana, and S. Shakkottai, "Joint scheduling of URLLC and eMBB traffic in 5G wireless networks," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Apr. 2018, pp. 1970–1978.

[2] J. Krause, *Study on Scenarios and Requirements for Next Generation Access Technology*, document TR 38.913, 3GPP, Sep. 2016.

[3] S. D'Oro, F. Restuccia, T. Melodia, and S. Palazzo, "Low-complexity distributed radio access network slicing: Algorithms and experimental results," *IEEE/ACM Trans. Netw.*, vol. 26, no. 6, pp. 2815–2828, Dec. 2018.

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

[5] Y. Guo, Q. Yang, and K. S. Kwak, "Quality-oriented rate control and resource allocation in time-varying OFDMA networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 3, pp. 2324–2338, Mar. 2017.

[6] P. Caballero, A. Banchs, G. De Veciana, and X. Costa-Perez, "Network slicing games: Enabling customization in multi-tenant mobile networks," *IEEE/ACM Trans. Netw.*, vol. 27, no. 2, pp. 662–675, Apr. 2019.

[7] P. Caballero, A. Banchs, G. de Veciana, and X. Costa-Pérez, "Multi-tenant radio access network slicing: Statistical multiplexing of spatial loads," *IEEE/ACM Trans. Netw.*, vol. 25, no. 5, pp. 3044–3058, Oct. 2017.

[8] A. A. Esswie and K. I. Pedersen, "Multi-user preemptive scheduling for critical low latency communications in 5G networks," in *Proc. IEEE Symp. Comput. Commun. (ISCC)*, Jun. 2018, pp. 136–141.

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

[10] X. Costa-Perez, J. Swetina, T. Guo, R. Mahindra, and S. Rangarajan, "Radio access network virtualization for future mobile carrier networks," *IEEE Commun. Mag.*, vol. 51, no. 7, pp. 27–35, Jul. 2013.

[11] S. H. da Mata and P. R. Guardieiro, "A genetic algorithm based approach for resource allocation in LTE uplink," in *Proc. Int. Telecommun. Symp. (ITS)*, Aug. 2014, pp. 1–5.

[12] J. Kwak, J. Moon, H.-W. Lee, and L. B. Le, "Dynamic network slicing and resource allocation for heterogeneous wireless services," in *Proc. IEEE 28th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Oct. 2017, pp. 1–5.

[13] Q. Ye, W. Zhuang, S. Zhang, A.-L. Jin, X. Shen, and X. Li, "Dynamic radio resource slicing for a two-tier heterogeneous wireless network," *IEEE Trans. Veh. Technol.*, vol. 67, no. 10, pp. 9896–9910, Oct. 2018.

[14] Y. Yao, L. Huang, A. Sharma, L. Golubchik, and M. Neely, "Data centers power reduction: A two time scale approach for delay tolerant workloads," in *Proc. Proc. IEEE INFOCOM*, Mar. 2012, pp. 1431–1439.

[15] B. Chatras, U. S. Tsang Kwong, and N. Bihannic, "NFV enabling network slicing for 5G," in *Proc. 20th Conf. Innov. Clouds, Internet Netw. (ICIN)*, Mar. 2017, pp. 219–225.

[16] M. Richart, J. Baliosian, J. Serrat, and J.-L. Gorricho, "Resource slicing in virtual wireless networks: A survey," *IEEE Trans. Netw. Service Manage.*, vol. 13, no. 3, pp. 462–476, Sep. 2016.

[17] J. S. Panchal, R. D. Yates, and M. M. Buddhikot, "Mobile network resource sharing options: Performance comparisons," *IEEE Trans. Wireless Commun.*, vol. 12, no. 9, pp. 4470–4482, Sep. 2013.

[18] F. Fu and U. C. Kozat, "Wireless network virtualization as a sequential auction game," in *Proc. Proc. IEEE INFOCOM*, Mar. 2010, pp. 1–9.

[19] H. Wang, C. Liu, L. Shen, and W. Xia, "Delay-aware resource allocation scheme for heterogeneous multi-radio access system based on Lyapunov optimization," in *Proc. 10th Int. Conf. Commun. Netw. China (ChinaCom)*, Aug. 2015, pp. 32–36.

[20] V. Sciancalepore, K. Samdanis, X. Costa-Perez, D. Bega, M. Gramaglia, and A. Banchs, "Mobile traffic forecasting for maximizing 5G network slicing resource utilization," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, May 2017, pp. 1–9.

[21] Y. Loong Lee, J. Loo, and T. Chee Chuah, "A new network slicing framework for multi-tenant heterogeneous cloud radio access networks," in *Proc. Int. Conf. Adv. Electr., Electron. Syst. Eng. (ICAEES)*, Nov. 2016, pp. 414–420.

[22] N. Merayo, P. Pavon-Marino, J. C. Aguado, R. J. Durán, F. Burrull, and V. Bueno-Delgado, "Fair bandwidth allocation algorithm for PONs based on network utility maximization," *J. Opt. Commun. Netw.*, vol. 9, no. 1, pp. 75–86, Jan. 2017.

[23] Y. Mao, J. Zhang, S. H. Song, and K. B. Letaief, "Power-delay tradeoff in multi-user mobile-edge computing systems," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2016, pp. 1–6.

[24] W. Chen, X. Xu, C. Yuan, J. Liu, and X. Tao, "Virtualized radio resource pre-allocation for QoS based resource efficiency in mobile networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2017, pp. 1–6.

[25] B. Gu, J. Feng, Z. Zhou, and M. Guizani, "Time-dependent pricing for on-demand bandwidth slicing in software defined networks," in *Proc. 14th Int. Wireless Commun. Mobile Comput. Conf. (IWCMC)*, Jun. 2018, pp. 1024–1029.

[26] P. Mogensen, W. Na, I. Z. Kovacs, F. Frederiksen, A. Pokhariyal, K. I. Pedersen, T. Kolding, K. Hugl, and M. Kuusela, "LTE capacity compared to the Shannon bound," in *Proc. IEEE 65th Veh. Technol. Conf. (VTC-Spring)*, Apr. 2007, pp. 1234–1238.

[27] G. Durisi, T. Koch, and P. Popovski, "Toward massive, ultrareliable, and low-latency wireless communication with short packets," *Proc. IEEE*, vol. 104, no. 9, pp. 1711–1726, Sep. 2016.

[28] P. Gao, S. Wittevrongel, K. Laevens, D. De Vleeschauwer, and H. Bruneel, "Distributional Little's law for queues with heterogeneous server interruptions," *Electron. Lett.*, vol. 46, no. 11, pp. 763–764, 2010.

[29] C.-F. Liu, M. Bennis, and H. V. Poor, "Latency and reliability-aware task offloading and resource allocation for mobile edge computing," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2017, pp. 1–7.

[30] M. Bennis, M. Debbah, and H. V. Poor, "Ultra-reliable and low-latency wireless communication: Tail, risk and scale," *Proc. IEEE*, vol. 106, no. 10, pp. 1834–1853, 2018.

[31] L. Tang, Y. Wei, L. He, H. Liao, and Q. Chen, "Queue-aware dynamic resource reuse and joint allocation algorithm in AC small cell networks," *IEEE Access*, vol. 6, pp. 61077–61090, 2018.

[32] Y. Wang, W. Wang, V. K. N. Lau, L. Chen, and Z. Zhang, "Heterogeneous spectrum aggregation: Coexistence from a queue stability perspective," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2471–2485, Apr. 2018.

[33] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6626–6637.

[34] Y. Yao, L. Huang, A. B. Sharma, L. Golubchik, and M. J. Neely, "Power cost reduction in distributed data centers: A two-time-Scale approach for delay tolerant workloads," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 1, pp. 200–211, Jan. 2014.

[35] M. Chen, M. Ponec, S. Sengupta, J. Li, and P. A. Chou, "Utility maximization in peer-to-peer systems with applications to video conferencing," *IEEE/ACM Trans. Netw.*, vol. 20, no. 6, pp. 1681–1694, Dec. 2012.

[36] Z. Wang, V. Aggarwal, and X. Wang, "Joint energy-bandwidth allocation in multiple broadcast channels with energy harvesting," *IEEE Trans. Commun.*, vol. 63, no. 10, pp. 3842–3855, Oct. 2015.

[37] J. Kwak, O. Choi, S. Chong, and P. Mohapatra, "Processor-network speed scaling for Energy–Delay tradeoff in smartphone applications," *IEEE/ACM Trans. Netw.*, vol. 24, no. 3, pp. 1647–1660, Jun. 2016.

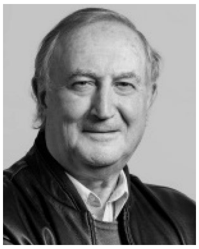
LEI FENG received the B.Eng. and Ph.D. degrees in communication and information systems from the Beijing University of Posts and Telecommunications (BUPT), in 2009 and 2015, respectively. He is currently a Lecturer with the State Key Laboratory of Networking and Switching Technology, BUPT. His research interests are resources management in wireless networks and smart grid.

YUEQI ZI received the B.Eng. and B.B.A. dual degrees from Queen Marry and Beijing University of Posts and Telecommunications (BUPT), respectively, in 2017, where she is currently pursuing the M.Eng. degree in computer science and technology. Her research interests are resources management in 5G wireless networks and smart grid.

WENJING LI is currently a Professor with BUPT and serves as the Director with the Key Laboratory of Network Management Research Center. Meanwhile, she is the Leader of TC7/WG1 in the China Communications Standards Association (CCSA). Her research interests are wireless network management and automatic healing in SONs.

FANQING ZHOU received the B.Eng. degree from the Beijing University of Posts and Telecommunications (BUPT), in 2012, where he is currently pursuing the Ph.D. degree in computer science and technology. His research interests include resources management and load balancing in multi-RAT heterogeneous networks.

PENG YU received the B.Eng. and Ph.D. degrees from BUPT, in 2008 and 2013, respectively. He is currently an Associate Professor with the State Key Laboratory of Networking and Switching Technology, BUPT. His research interests are autonomic management and hybrid energy allocation in GreenNet.



MICHEL KADOCH received the B.Eng. degree from Sir George Williams University, in 1971, the M.Eng. degree from Carleton University, in 1974, the M.B.A. degree from McGill University, in 1983, and the Ph.D. degree from Concordia University, in 1991.

He is currently a Full Professor with Ecole de technologie superieure (ETS), Universite du Quebec, Montreal, QC, Canada. He is also an Adjunct Professor with Concordia University, Montreal. He is serving as a Reviewer for a number of journals and conferences, as well as for NSERC grants. His current research interests include crosslayer design, reliable multicast in wireless *ad hoc*, and WiMAX networks. He has publications and patents in all these areas.