Delay-Aware Resource Allocation for Network Slicing in Fog Radio Access Networks

Dongxue Tang, Chunjing Hu, and Tian Dang
Key Laboratory of Universal Wireless Communication, Ministry of Education
Beijing University of Posts and Telecommunications, Beijing, 100876, China
Email:tdx24601@tju.edu.cn

***Abstract*—To meet diverse quality of service (QoS) requirements in fifth generation (5G) wireless networks, network slicing is proposed as a cost-efficient way to allow operators to provide customized services. On the other hand, fog radio access network (F-RAN) has emerged as a promising network architecture to satisfy different QoS requirements in distinct application scenarios. To make the network slicing work more efficiently, a downlink F-RAN, equipped with a ultra-reliable and low-latency communication (URLLC) slice and an enhanced mobile broadband (eMBB) slice logically, is concerned in this paper. A delay-aware resource allocation optimization problem is formulated, aiming to maximize the throughput of the eMBB slice as well as guarantee the queuing delay performance of the URLLC slice. The delay-aware optimization problem is reformulated as a per-slot joint resource block (RB) and power allocation problem, which is addressed by Lagrange dual decomposition method. Simulation results show a tradeoff between the throughput and the queuing delay of eMBB slice can be achieved while satisfying the delay constraint of URLLC slice.**

I. Introduction

In order to satisfy diverse QoS requirements of different applications in fifth generation (5G) wireless networks, network slicing has been proposed as a cost-efficient way to allow operators to provide customized services. For network slicing [1], the physical network is divided into multiple logically independent networks as network slices, and each slice has its corresponding network functions and radio access technique. Meanwhile, fog radio access networks (F-RAN) [2] has emerged as a promising network architecture that be adaptive to distinct application scenarios [3]. Furthermore, traditional core network (CN)-based network slicing cannot fully exploit the potential of F-RAN, investigation of network slicing in F-RAN is necessary.

Radio resource management (RRM) is critical to the performance of network slicing, and many recent works have investigated the resource management in network slicing. In [4], a joint admission control and resource optimization algorithm was proposed to maximize the sum rate of a single-cell wireless virtualized network (WVN) while satisfying the minimum rate constraint of rate-based slices and the minimum power and sub-carriers constraint of resource-based slices. In [5], a distributed algorithm was proposed to maximize the revenue earned by mobile virtual network operator under satisfying users' minimum rate requirement and backhaul link constraint. Considering the ever-growing demand of QoS requirement of latency and reliability imposed by the URLLC communications, delay-aware resource optimization has been studied in works [6] and [7]. In [6], the joint optimization of power allocation and content matching is investigated for wireless content caching networks to maximize the transmission success probability which implies the reliability of networks. A stochastic optimization problem is studied in [7] to maximize the overall throughput of C-RANs with device-to-device communications.

Due to the fact that there are very few solutions considering different QoS requirements of slices in multi-types application scenario, in this paper we propose a joint resource block (RB) and power allocation scheme which considers not only the channel state information (CSI) but also the queue state information and different application specific QoS requirements. We formulate a joint RB and power allocation problem in the downlink of F-RANs with URLLC user's data queue length probabilistic reliability constraints, RB assignment and power allocation constraints to maximize the throughput of eMBB slice as well as satisfy the latency and reliability requirement of URLLC slice. With Lyapunov optimization framework, the primal mixed integer non-linear programming problem is transformed and then addressed by Lagrange dual decomposition method. Simulations results show that the proposed resource allocation algorithm which has taken into account the interference in URLLC slice can achieve a tradeoff between the throughput and the queueing delay of eMBB slice while satisfying the delay performance requirement of URLLC slice.

The rest of this paper is organized as follows. Section II presents the system model and formulates optimization problem. In section III, the primal problem is transformed and solved. The simulation results and analysis are given in Section IV, followed by the conclusion in Section V.

II. System Model and Problem Formulation

A. System Model

As shown in Fig. 1, a downlink F-RAN system consisting of one eMBB slice and one URLLC slice is investigated. The former provides enhanced mobile broadband services while the latter caters to ultra-reliable and low latency services. The eMBB slice is configured with N single-antenna RRHs, each user of which is served by all RRHs with global mode. The URLLC slice is consisted of D D2D pairs and one gNodeB (gNB) with cached content serving M user equipments (UEs).

978-1-5386-6119-2/18/$31.00 ©2018 IEEE

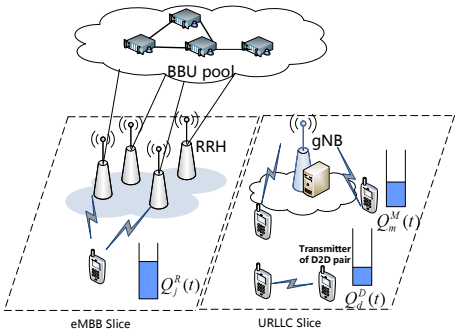

Fig. 1. Downlink F-RANs with two network slices for different services

The UEs served by RRHs are denoted by RUEs, the UEs served by gNB are labeled as GUEs, and UEs operating on D2D model are labeled by DUEs. Let $\mathcal{N} = \{1, 2, ..., N\}$ represent the set of RRHs, and $\mathcal{J} = \{1, 2, ..., J\}$ denote the set of RUEs. Let $\mathcal{M} = \{1, 2, ..., M\}$ denote the set of GUEs, and $\mathcal{D} = \{1, 2, ..., D\}$ denote the set of D2D pairs. The bandwidth of each RB is denoted by W_0, and the entire system bandwidth is W. The RBs assigned to different slices are orthogonal, therefore it is inter-slice interference-free between eMBB slice and URLLC slice. Specially, the RBs assigned to different RUEs in eMBB slice are orthogonal, while the RBs in URLLC slice are reused by one GUE and one D2D pair in URLLC slice, therefore it is intra-tier interference-free in eMBB slice and there is inter-tier interference in URLLC slice. Note that the additive white Gaussian noise is distributed as $\mathcal{CN}(0, \sigma^2)$.

The network operates in slotted time $t \in \{1, 2, ..., T\}$ with length of each time slot denoted by $\tau = 1$. Let h_{ijk} , g_{mk} , g_{ddk} denote the CSI on RB k from RRH i to RUE j, the gNB to GUE m and the transmitter to the receiver in D2D pair d. Let p^R_{ijk} represent the transmit power from RRH i to RUE j on RB k , p^G_{mk} denote the transmit power from the gNB to GUE m on RB k, and p^D_{dk} the transmit power from transmitter to receiver in D2D pair d on RB k. In addition, the binary variables a_{jk}, a_{mk} and a_{dk} represent the allocation of RB k to RUE j, the allocation of RB k to GUE m, and to D2D pair d, which satisfy the following constraints for all k and t:

$$\text{C1}: \sum_{j=1}^{J} a_{jk}(t) \le 1, \sum_{m=1}^{M} a_{mk}(t) \le 1, \sum_{d=1}^{D} a_{dk}(t) \le 1, \quad (1)$$

$$\text{C2}: \sum_{m=1}^{M} a_{mk}(t) + \sum_{d=1}^{D} a_{dk}(t) \in \{0, 2\}, \quad (2)$$

$$\text{C3}: \sum_{j=1}^{J} a_{jk}(t) + \varphi\left\{\sum_{m=1}^{M} a_{mk}(t) + \sum_{d=1}^{D} a_{dk}(t)\right\} = 1, \quad (3)$$

where $\varphi(x) = 1,\ if\ x > 0$ and $\varphi(x) = 0,\ if\ x = 0$. The inter-tier interference on RB k from the transmitters of D2D pairs to GUE m in URLLC slice is denoted by I_{mk}, and interference on RB k from GUEs to the receiver of D2D pair d is expressed as I_{dk}, which can be given by

$$I_{mk}(t) = \sum_{d=1}^{D} a_{dk}(t)\, p^D_{dk}(t)\, |g_{dmk}(t)|^2, \quad (4)$$

$$I_{dk}(t) = \sum_{m=1}^{M} a_{mk}(t)\, p^G_{mk}(t)\, |g_{dk}(t)|^2, \quad (5)$$

where g_{dmk} denotes the CSI on RB k from the transmitter of D2D pair d to the GUE m, and g_{dk} denotes the CSI on RB k from gNB to the receiver of D2D pair d. Therefore the downlink transmission rate for RUE j, GUE m, and the receiver of D2D pair d can be given by

$$R^{\mathrm{R}}_j(t) = \sum_{k=1}^{K} a_{jk}(t)\, W_0 \log_2\Big(1 + \frac{\sum_{i=1}^{N} |h_{ijk}(t)|^2 p^R_{ijk}(t)}{\sigma^2}\Big), \quad (6)$$

$$R^{\mathrm{G}}_m(t) = \sum_{k=1}^{K} a_{mk}(t)\, W_0 \log_2\Big(1 + \frac{|g_{mk}(t)|^2 p^G_{mk}(t)}{\sigma^2 + I^{G\max}_{mk}}\Big), \quad (7)$$

$$R^{\mathrm{D}}_d(t) = \sum_{k=1}^{K} a_{dk}(t)\, W_0 \log_2\Big(1 + \frac{|g_{ddk}(t)|^2 p^D_{dk}(t)}{\sigma^2 + I^{D\max}_{dk}}\Big). \quad (8)$$

For simplification, in (7) and (8), we replace the inter-tier interference term in the denominator of logarithmic expression of rate with the maximum interference $I^{D\max}_{dk}$ that the receiver of D2D pair d can tolerlate on RB k and the maximum interference $I^{G\max}_{mk}$ that GUE m can tolerlate on RB k, which satisfy the following constraints:

$$\text{C4}: \quad a_{dk}(t)\, I_{dk}(t) \le a_{dk}(t)\, I^{D\max}_{dk}, \forall d, k \quad (9)$$

$$\text{C5}: \quad a_{mk}(t)\, I_{mk}(t) \le a_{mk}(t)\, I^{G\max}_{mk}, \forall m, k \quad (10)$$

which is proved to be effective in [8]. The transmit power of RRH i, gNB and the transmitter of D2D pair d is given by

$$p_i{}^R(t) = \sum_{j=1}^{J}\sum_{k=1}^{K} a_{jk}(t)\, p^R_{ijk}(t), \quad (11)$$

$$p^G_g(t) = \sum_{m=1}^{M}\sum_{k=1}^{K} a_{mk}(t)\, p^G_{mk}(t), \quad (12)$$

$$p^D_d(t) = \sum_{k=1}^{K} a_{dk}(t)\, p^D_{dk}(t). \quad (13)$$

Denote buffering queue length of RUE j, GUE m, and the transmitter of D2D pair d as $Q^R_j(t)$, $Q^G_m(t)$, and $Q^D_d(t)$ respectively. Let $A^R_j(t)$, $A^G_m(t)$ and $A^D_d(t)$ represent the traffic arrival of RUE j, GUE m, the transmitter of D2D pair d at slot t. The evolutions of $Q^R_j(t)$, $Q^G_m(t)$, and $Q^D_d(t)$ are given by

$$Q_j^R(t+1) = \left\{Q_j^R(t) - R_j^R(t)\right\}^+ + A_j^R(t), \quad (14)$$

$$Q_m^G(t+1) = \left\{Q_m^G(t) - R_m^G(t)\right\}^+ + A_m^G(t), \quad (15)$$

$$Q_d^D(t+1) = \left\{Q_d^D(t) - R_d^D(t)\right\}^+ + A_d^D(t), \quad (16)$$

where $\{x\}^+ \triangleq \max\{x, 0\}$. According to Little's theorem [9], the average delay is proportional to the average queue length. Note that the violation of queue length bound can be regarded as a destruction to reliability, thus a probabilistic constraint of queue length is imposed on each GUE m and each D2D pair d, i.e., $\Pr(Q_m^G(t) \geq L_m) \leq \xi_m$, and $\Pr(Q_d^D(t) \geq L_d) \leq \xi_d$. L_m and L_d are the allowable maximum queue lengths of GUE m and D2D pair d, while the tolerance values ξ_m and ξ_d represent reliability.

B. Problem Formulation

Let $\mathbf{p} = [p_{ijk}(t), p_{mk}(t), p_{dk}(t) : i \in \mathcal{N}, j \in \mathcal{J}, m \in \mathcal{M}, d \in \mathcal{D}, k \in \mathcal{K}]$ and $\mathbf{a} = [a_{jk}(t), a_{mk}(t), a_{dk}(t) : j \in \mathcal{J}, m \in \mathcal{M}, d \in \mathcal{D}, k \in \mathcal{K}]$ denote the vectors of power allocation and RB assignment respectively. The optimization objective is to maximize the time-average throughput of eMBB slice while satisfying the delay constraints of URLLC slice. With the consideration of RB and power allocation constraints, the stochastic optimization problem is formulated as

$$\begin{aligned}
&\max_{\{\mathbf{p},\mathbf{a}\}} \sum_{j=1}^{J} \bar{R}_j^R(t) \\
&\text{s.t.} \quad \text{C1} - \text{C5} \\
&\text{C6}: p_i^R(t) \leq {p_i}^{R\max}, \forall i, t, \\
&\text{C7}: p_g^G(t) \leq {p_g}^{G\max}, \forall t, \\
&\text{C8}: \ p_d^D(t) \leq {p_d}^{D\max}, \forall d, t, \\
&\text{C9}: \ \lim_{t\to\infty} \frac{\mathrm{E}\{|Q_j^R(t)|\}}{t} = 0, \forall j, \\
&\text{C10}: \ \lim_{t\to\infty} \Pr\left(Q_m^G(t) \geq L_m\right) \leq \xi_m, \\
&\qquad \lim_{t\to\infty} \Pr(Q_d^D(t) \geq L_d) \leq \xi_d, \forall m, d,
\end{aligned} \quad (17)$$

Constraint C1 ensures that each RB can be assigned to no more than one user of the same type, while C2 ensures that each RB assigned to URLLC slice is reused by one GUE and one D2D pair. C3 ensures that each RB cannot be assigned to more than one slice. C4 and C5 restrict the maximum tolerable interference level of the receiver of D2D pair d and GUE m on RB k if RB k is assigned to D2D pair d and GUE m respectively. C6-C8 restrict the transmit power of RRHs, gNB and transmitters of D2D pairs. C9 is the queue stability constraint for RUEs, C10 is the probabilistic constraint on queue length of GUEs and transmitters of D2D pairs. Note that the problem in (17) is a mixed integer programming problem which is non-convex and NP-hard in general.

III. RESOURCE BLOCK AND POWER ALLOCATION

The optimization problem has a probabilistic constraint C10, which is untractable to solve. To handle the constraint, the Markov's inequality [10] is used to transform C10 such that $\Pr\{Q_m^G(t) \geq L_m\} \leq \mathbb{E}\{Q_m^G(t)\}/L_m$. Then C10 is satisfied if

$$\bar{Q}_m^G = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left\{Q_m^G(t)\right\} \leq L_m \xi_m, \forall m, \quad (18)$$

$$\bar{Q}_d^D = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left\{Q_d^D(t)\right\} \leq L_d \xi_d, \forall d, \quad (19)$$

(18) and (19) will be satisfied if the constructed virtual queues are mean-rate stable, whose evolutions are given by

$$F_m^G(t+1) = \max\left\{F_m^G(t) + Q_m^G(t+1) - L_m \varepsilon_m, 0\right\} \quad (20)$$

$$F_d^D(t+1) = \max\left\{F_d^D(t) + Q_d^D(t+1) - L_d \varepsilon_d, 0\right\} \quad (21)$$

With the definitions of different queues, $\mathbf{\Theta}(t) = \left[\mathbf{Q}^R(t), \mathbf{F}^G(t), \mathbf{F}^D(t)\right]$ represents the combined matrix of partial actual queues and all virtual queues, where $\mathbf{Q}^R(t) = \left\{Q_j^R(t) | j \in \mathcal{J}\right\}$, $\mathbf{F}^G(t) = \left\{F_m^G(t) | m \in \mathcal{M}\right\}$ and $\mathbf{F}^D(t) = \left\{F_d^D(t) | d \in \mathcal{D}\right\}$. The Lyapunov function is defined as

$$L(\mathbf{\Theta}(t)) \triangleq \frac{1}{2} \left\{ \sum_{j=1}^{J} Q_j^R(t)^2 + \sum_{m=1}^{M} F_m^G(t)^2 + \sum_{d=1}^{D} F_d^D(t)^2 \right\} \quad (22)$$

Then the conditional Lyapunov drift-plus-penalty [11] for slot t is given by

$$\Delta(\mathbf{\Theta}(t)) - V\mathbb{E}\left\{ \sum_{j=1}^{J} R_j^R(t) | \mathbf{\Theta}(t) \right\} \quad (23)$$

where $\Delta(\mathbf{\Theta}(t)) \triangleq \mathbb{E}\left\{L(\mathbf{\Theta}(t+1)) - L(\mathbf{\Theta}(t)) | \mathbf{\Theta}(t)\right\}$ which is the Lyapunov drift. V is a non-negative parameter handling the tradeoff between the throughput of eMBB slice and latency reduction. Substituting (14)-(16), (20) and (21) into (22), and using $(\max\{x, 0\})^2 \leq x^2$, then by (23), we have

$$\begin{aligned}
\Delta\big(\mathbf{\Theta}(t)\big) - V\mathbb{E}\left\{ \sum_{j=1}^{J} R_j^R(t) | \mathbf{\Theta}(t) \right\} \leq B - V \sum_{j=1}^{J} R_j^R(t) \\
- \sum_{j=1}^{J} \left[Q_j^R(t) + A_j^R(t)\right] E\left\{R_j^R(t)\right\} \\
- \sum_{m=1}^{M} \left[F_m^G(t) + Q_m^G(t) + A_m^G(t)\right] E\left\{R_m^G(t)\right\} \\
- \sum_{d=1}^{D} \left[F_d^D(t) + Q_d^D(t) + A_d^D(t)\right] E\left\{R_d^D(t)\right\}
\end{aligned} \quad (24)$$

where B is a constant. Since the constant has no impact on the problem, we omit its expression. With Lyapunov optimization framework [11], the primal problem is transformed into minimizing the right hand side of (24), given by

$$\begin{aligned}\max_{\{\mathbf{a},\mathbf{p}\}} &\sum_{j=1}^{J}\left[V+Q_j^R(t)+A_j^R(t)\right]R_j^R(t)\\ &+\sum_{m=1}^{M}\left[F_m^G(t)+Q_m^G(t)+A_m^G(t)\right]R_m^G(t)\\ &+\sum_{d=1}^{D}\left[F_d^D(t)+Q_d^D(t)+A_d^D(t)\right]R_d^D(t)\\ &\text{s.t.}\quad \text{C1}-\text{C8}.\end{aligned} \tag{25}$$

With continuity relaxation of binary variables, the optimization problem can be transformed to be convex and addressed by the Lagrange dual decomposition method [12]. Under constraints C4-C8, we attain the Lagrangian function of the transformed objective function as follows:

$$\begin{aligned}L(\mathbf{a},\mathbf{p},\beta,\lambda,\gamma,\eta,\nu)=&\sum_{j=1}^{J}\left[Q_j^R+A_j^R+V\right]R_j^R\\ &+\sum_{m=1}^{M}\left[F_m^G+Q_m^G+A_m^G\right]R_m^G\\ &+\sum_{d=1}^{D}\left[F_d^D+Q_d^D+A_d^D\right]R_d^D\\ &+\sum_{i=1}^{N}\beta_i({p_i}^{R\max}-\sum_{j=1}^{J}\sum_{k=1}^{K}a_{jk}p_{ijk}^R)\\ &+\nu({p_g}^{G\max}-\sum_{m=1}^{M}\sum_{k=1}^{K}a_{mk}p_{mk}^G)\\ &+\sum_{d=1}^{D}\lambda_d({p_d}^{D\max}-\sum_{k=1}^{K}a_{dk}p_{dk}^D)\\ &+\sum_{k=1}^{K}\sum_{m=1}^{M}\gamma_{m,k}a_{m,k}[I_{mk}^{G\max}-\sum_{d=1}^{D}a_{dk}p_{dk}^D|g_{dmk}|^2]\\ &+\sum_{k=1}^{K}\sum_{d=1}^{D}\eta_{d,k}a_{d,k}[I_{dk}^{D\max}-\sum_{m=1}^{M}a_{mk}p_{mk}^G|g_{dk}|^2]\end{aligned} \tag{26}$$

where $\beta=(\beta_1,\beta_2,...,\beta_N)\succeq 0, \lambda=(\lambda_1,\lambda_2,...,\lambda_D)\succeq 0$, $\nu \geq 0$ are the Lagrange multiplier vectors related to transmit power constraints of RRHs, gNB and transmitters of D2D pairs. $\gamma=(\gamma_{1,1},\gamma_{1,2},...,\gamma_{1,k},\gamma_{2,1},\gamma_{2,2},...,\gamma_{2,k},...,\gamma_{m,k})\succeq 0$ and $\eta=(\eta_{1,1},\eta_{1,2},...,\eta_{1,k},\eta_{2,1},\eta_{2,2},...,\eta_{2,k},...,\eta_{d,k})\succeq 0$ are the Lagrange multiplier vectors associated with the inter-tier interference constraints in URLLC slice.

The Lagrangian dual function is given by

$$\begin{aligned}g(\beta,\lambda,\gamma,\eta,\nu)=&\max_{\{\mathbf{a},\mathbf{p}\}}L(\mathbf{a},\mathbf{p},\beta,\lambda,\gamma,\eta,\nu)\\ &\text{s.t.}\quad \text{C1}-\text{C3}\end{aligned} \tag{27}$$

The dual optimization problem can be expressed as

$$\begin{aligned}&\min_{\{\beta,\lambda,\gamma,\eta,\nu\}} g(\beta,\lambda,\gamma,\eta,\nu)\\ &\text{s.t.}\quad \beta\succeq 0,\lambda\succeq 0,\gamma\succeq 0,\eta\succeq 0,\nu\geq 0\end{aligned} \tag{28}$$

Note that the Lagrangian function $L(\mathbf{a},\mathbf{p},\ \beta,\lambda,\gamma,\eta,\nu)$ is linear with β_i, λ_d, ν, $\gamma_{m,k}$, $\eta_{d,k}$ for any fixed a_{jk}, a_{dk}, a_{mk}. Then the dual decomposition method is applied to decompose the problem into K independent sub-problems as

$$\begin{aligned}g(\beta,\lambda,\gamma,\eta,\nu)=&\sum_{k=1}^{K}g_k(\beta,\lambda,\gamma,\eta,\nu)\\ &+\sum_{i=1}^{N}\beta_i{p_i}^{R\max}+\sum_{d=1}^{D}\lambda_d{p_d}^{D\max}+\nu{p_g}^{G\max}\end{aligned} \tag{29}$$

where

$$\begin{aligned}g_k(\beta,\lambda,\gamma,\eta,\nu)=&\max_{\{\mathbf{a},\mathbf{p}\}}\\ &\Bigg\{\sum_{j=1}^{J}a_{jk}W_0\left[Q_j^R+A_j^R+V\right]\log_2\Big(1+\frac{\sum_{i\in R}|h_{ijk}|^2p_{ijk}^R}{\sigma^2}\Big)\\ &+\sum_{m=1}^{M}a_{mk}W_0\left[F_m^G+Q_m^G+A_m^G\right]\log_2\Big(1+\frac{|g_{mk}|^2p_{mk}^G}{\sigma^2+I_{mk}^{G\max}}\Big)\\ &+\sum_{d=1}^{D}a_{dk}W_0\left[F_d^D+Q_d^D+A_d^D\right]\log_2\Big(1+\frac{|g_{ddk}|^2p_{dk}^D}{\sigma^2+I_{dk}^{D\max}}\Big)\\ &-\sum_{i=1}^{N}\sum_{j=1}^{J}\beta_ia_{jk}p_{ijk}^R-\sum_{d=1}^{D}\lambda_da_{dk}p_{dk}^D-\sum_{m=1}^{M}\nu a_{mk}p_{mk}^G\\ &+\sum_{m=1}^{M}\gamma_{m,k}a_{m,k}\Big[I_{mk}^{G\max}-\sum_{d=1}^{D}a_{dk}p_{dk}^D|g_{dmk}|^2\Big]\\ &+\sum_{d=1}^{D}\eta_{d,k}a_{d,k}\Big[I_{dk}^{D\max}-\sum_{m=1}^{M}a_{mk}p_{mk}^G|g_{dk}|^2\Big]\Bigg\}\end{aligned} \tag{30}$$

Suppose that the kth RB is assigned to RUE j, i.e. $a_{jk}=1$, it is observed that (30) is concave with respect to p_{ijk}^R. According to the KKT condition, the optimal power allocation can be given by

$$p_{ijk}^{R,*}=\Big\{\frac{W_0\left[Q_j^R+A_j^R+V\right]}{\beta_i\ln 2}-\frac{\sigma^2+\sum_{i'\neq i}|h_{i'jk}|^2p_{i'jk}^{R,*}}{|h_{ijk}|^2}\Big\}^+. \tag{31}$$

Suppose that the kth RB is assigned to GUE m and D2D pair d, i.e. $(a_{mk},a_{dk})=(1,1)$, it is noted that (30) is concave with respect to p_{mk}^G, p_{dk}^D respectively. The optimal power

allocation can be derived by

$$p_{mk}^{G,*} = \left\{ \frac{W_0 \left[F_m^G + Q_m^G + A_m^G\right]}{\left(\nu + \eta_{d,k}|g_{dk}|^2\right) \ln 2} - \frac{\sigma^2 + I_{mk}^{G\max}}{|g_{mk}|^2} \right\}^+ \quad (32)$$

$$p_{dk}^{D*} = \left\{ \frac{W_0 \left[F_d^D + Q_d^D + A_d^D\right]}{\left(\lambda_d + \gamma_{m,k}|g_{dmk}|^2\right) \ln 2} - \frac{\sigma^2 + I_{dk}^{D\max}}{|g_{ddk}|^2} \right\}^+ \quad (33)$$

Substitute the optimal power allocations $p_{ijk}^{R,*}$, $p_{mk}^{G,*}$, and p_{dk}^{D*} into (30), and denote

$$\begin{aligned}
\phi_{jk} =& W_0 \left[Q_j^R + A_j^R + V\right] \log_2\left(1 + \frac{\sum\limits_{i\in R} |h_{ijk}|^2 p_{ijk}^R}{\sigma^2}\right) \\
& - \sum_{i=1}^{N} \beta_i p_{ijk}^R \\
\Lambda_{mk} =& W_0 \left[F_m^G + Q_m^G + A_m^G\right] \log_2\left(1 + \frac{|g_{mk}|^2 p_{mk}^G}{\sigma^2 + I_{mk}^{G\max}}\right) \\
& - \nu p_{mk}^G + \gamma_{m,k}\left(I_{mk}^{G\max} - p_{dk}^D |g_{dmk}|^2\right) \\
\Gamma_{dk} =& W_0 \left[F_d^D + Q_d^D + A_d^D\right] \log_2\left(1 + \frac{|g_{ddk}|^2 p_{dk}^D}{\sigma^2 + I_{dk}^{D\max}}\right) \\
& - \lambda_d p_{dk}^D + \eta_{d,k}\left[I_{dk}^{D\max} - p_{mk}^G |g_{dk}|^2\right] \quad (34)
\end{aligned}$$

Then the dual function can be simplified into

$$\begin{aligned}
&\max_{\mathbf{a}} \sum_{k=1}^{K}\sum_{j=1}^{J} a_{jk}\phi_{jk} + \sum_{k=1}^{K}\sum_{m=1}^{M} a_{mk}\Lambda_{mk} + \sum_{k=1}^{K}\sum_{d=1}^{D} a_{dk}\Gamma_{dk} \\
&\text{s.t.} \quad \text{C1} - \text{C3} \\
&a_{jk}, a_{mk}, a_{dk} \in [0,1], \forall j, m, d, k, t
\end{aligned}$$

which is a linear integer programming problem. The optimal RB assignment can be attained by the following scheme with use of continuity relaxation. For the RB k, the RB assignment to RUE j is according to

$$a_{jk} = \begin{cases} 1, \text{ if } j = \text{argmax}\{\phi_{jk} : 1 \le j \le J\} \\ \quad \&\phi_{jk} > \max\{\Lambda_{mk} + \Gamma_{dk} : (m,d), \\ \quad 1 \le m \le M, 1 \le d \le D\} \\ 0, \text{ otherwise.} \end{cases} \quad (35)$$

For the RB k, the RB assignment to GUE j and D2D pair d is according to

$$(a_{mk}, a_{dk}) = \begin{cases} (1,1), \text{ if } (m,d) = \arg\max\{\Lambda_{mk} + \Gamma_{dk} : \\ \quad (m,d), 1 \le m \le M, 1 \le d \le D\} \& \\ \quad \Lambda_{mk} + \Gamma_{dk} > \max\{\phi_{jk} : 1 \le j \le J\} \\ (0,0), \text{ otherwise.} \end{cases} \quad (36)$$

To derive the optimal primal solution, we compute the dual variables iteratively with the subgradient method. Due to space restrictions, the expressions of the subgradient of the dual function and the expressions for updating the dual variables are omitted. Finally, the overall procedure of RB assignment and power allocation is described in Algorithm 1.

Algorithm 1 RB Assignment and Power Allocation

1: In each time slot, observe the actual queues $Q_j^R(t)$, $Q_m^G(t)$, $Q_d^D(t)$ and the virtual queues $F_m^G(t)$, $F_d^D(t)$;
2: **repeat**
3: Compute the optimal power allocation of p_{ijk}^R by iteratively updating (31);
4: Compute the optimal power allocation of gNB p_{mk}^G and D2D pair p_{dk}^D according to (32) and (33);
5: Compute the optimal RB assignment a_{jk}, a_{mk}, a_{dk} according to (35) and (36);
6: Updating the Lagrangian dual variables $\beta, \lambda, \gamma, \eta, \nu$;
7: **until** Convergence;
8: Update the traffic queues $Q_j^R(t)$, $Q_m^G(t)$, $Q_d^D(t)$ and the virtual queues $F_m^G(t)$, $F_d^D(t)$ according to (14)-(16), (20) and (21).

IV. Simulation Results And Analysis

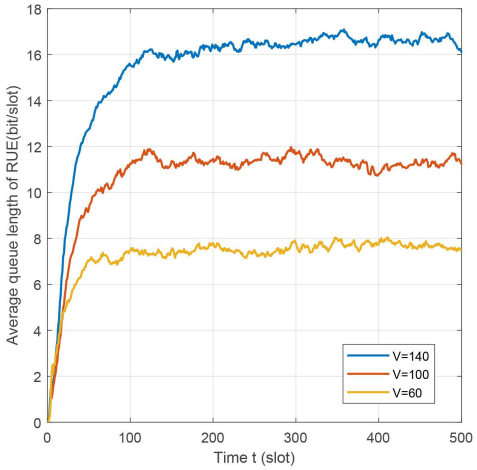

Fig. 2. Queue length of RUEs versus time slot t.

A downlink system of F-RANs configured with two slices is considered in simulations. The eMBB slice is configured with two single-antenna RRHs and 5 RUEs, while the URLLC slice is consist of 3 D2D pairs and one gNB serving 3 GUEs. The number of RBs is $K = 20$ and the bandwidth of each RB is $W_0 = 15$ kHz. We set ${p_i}^{R\max} = 26$ dBm, ${p_g}^{G\max} = 30$ dBm, and ${p_d}^{D\max} = 23$ dBm. The noise power spectrum density σ^2 is -174 dBm. we assume the traffic arrivals of RUEs, GUEs and the transmitter of D2D pairs to be constant in each time slots without loss of generality. The mean traffic arrival rates are given by $\bar{A}_j^R = 1.0$ bit/slot/Hz, $\bar{A}_m^G = 0.6$ bit/slot/Hz and $\bar{A}_d^D = 0.4$ bit/slot/Hz. We set the allowable maximum queue length $L_m = 8.8$ bit, $L_d = 6.6$ bit and the tolerance maximum probability that queue length exceeds allowable maximum queue $\xi_m = 0.1$, $\xi_d = 0.1$.

In Fig. 2, we evaluate the average queue length of RUE over 500 time slots. It is shown that the queue length of

RUE grows with t and then floats within a certain range. Therefore, the queue stability constraint of RUE is satisfied. In addition, we evaluate stable queue lengths of RUE under different values of control parameter V. It is obvious that a larger V will lead to worse delay performance because of the fact that more emphasis is put on throughput of eMBB slice instead of queue length minimization. Moreover, by analyzing simulation results, the probability that queue lengths of GUE and DUE exceed the given allowable maximum queue lengths L_m, L_d are 0.17% and 0.13% respectively under the given control parameter $V = 100$, which is smaller than ξ_m and ξ_d. Thus, the delay performance requirement of URLLC slice is satisfied.

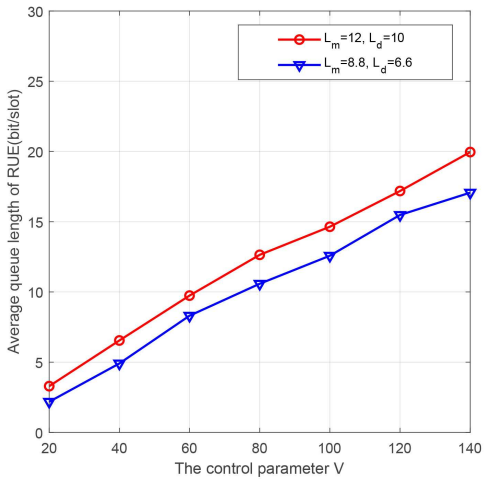

Fig. 3. Queue length of RUE versus control parameter V.

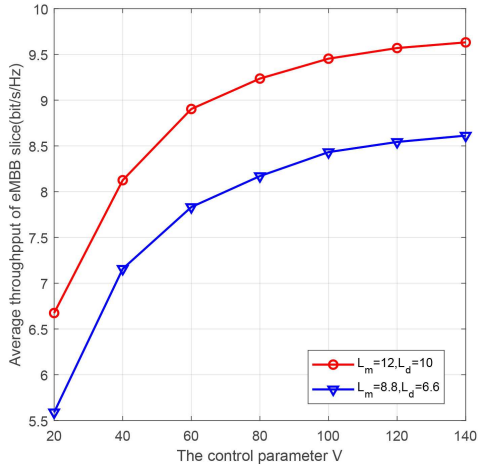

Fig. 4. Throughput of eMBB slice versus control parameter V.

Fig.3 presents the average queue length of RUE against the control parameter V. It is shown that the average queue length of RUE grows linearly at $O(V)$ with given average traffic arrival rate. Fig. 4 plots the overall throughput of eMBB slice versus different control parameter V. It can be observed that a larger V will lead to worse delay performance and greater throughput in eMBB slice because a larger V emphasizes more on the latter. For the same V, larger allowable maximum queue lengths of GUE and DUE lead to a better throughput performance because less RBs needed to be allocated to URLLC slice due to looser delay performance requirement of URLLC slice. Combining Fig.3 and Fig.4, it is clearly that there exists a tradeoff between the throughput and queuing delay of eMBB slice with a control parameter V for adjustment.

V. CONCLUSION

This paper has focused on a delay-aware resource allocation problem for enhanced mobile broadband (eMBB) and ultra-reliable and low-latency communication (URLLC) slices in the downlink of fog radio access networks (F-RAN). The optimization problem aimed to maximize the throughput of the eMBB slice as well as guarantee the queuing delay performance of the URLLC slice. The primal optimization problem was transformed with Lyapunov optimization framework and then addressed by Lagrange dual decomposition method. Simulation results showed that the proposed resource allocation algorithm attained a tradeoff between the throughput and queueing delay of eMBB slice with the probability that user queue length of URLLC slice exceeds the allowable maximum value under given threshold.

VI. ACKNOWLEDGEMENT

This work was supported by the National Natural Science Foundation of China under Grant No. 61671074.

REFERENCES

[1] R. Hattachi and J. Erfanian, "5G White Paper, " tech. rep., Feb. 2015.
[2] M. Peng and K. Zhang, "Recent advances in fog radio access networks: performance analysis and radio resource allocation," *IEEE Network*, vol. 4, pp. 5003-5009, Aug. 2016.
[3] M. Peng, S. Yan, K. Zhang, and C. Wang, "Fog-computing-based radio access networks : Issues and challenges," *IEEE Access*, vol. 30, no. 4, pp. 46-53, July-Aug. 2016.
[4] S. Parsaeefard, V. Jumba, M. Derakhshani, *et al.*, "Joint resource provisioning and admission control in wireless virtualized networks, " in *Wireless Communications and Networking Conference (WCNC), 2015 IEEE*, pp. 2020-2025, Mar. 2015.
[5] LeAnh T, Tran N H, Ngo D T, *et al.* "Resource allocation for virtualized wireless networks with backhaul constraints," *IEEE Commun. Lett.*, vol. 21, no. 1, pp. 148-151, Jan. 2016.
[6] Z. Zhao, M. Xu, Y. Li, and M. Peng, "A non-orthogonal multiple access-based multicast scheme in wireless content caching networks," *IEEE J. Sel. Areas Commun.*,vol. 35, no. 12, pp. 2723-2735, Dec. 2017.
[7] Y. Mo, M. Peng, H. Xiang, *et al.* "Resource allocation in cloud radio access networks with device-to-device communications," *IEEE Access*, vol. 5, pp. 1250-1262, Mar. 2017.
[8] A. Abdelnasser and E. Hossain, "Resource allocation for an OFDMA cloud-RAN of small cells underlaying a macrocell," *IEEE Trans. Mobile Comput.*, vol. 15, no. 11, pp. 2837-2850, Nov. 2016.
[9] J. D. Little and S. C. Graves, "Littles law," in *Building intuition.*Springer, pp. 81-100, 2008.
[10] A. Mukherjee, "Queue-aware dynamic on/off switching of small cells in dense heterogeneous networks," in *Proc. IEEE Global Commun. Conf. Workshops, Atlanta*, GA, USA, pp. 182-187, Dec. 2013.
[11] M. Neely, *Stochastic NetworkOptimizationWith Application toCommunication and Queuing Systems*. San Rafael, CA, USA: Morgan&Claypool, 2010.
[12] M. Peng, K. Zhang, J. Jiang, *et al.* "Energy-efficient resource assignment and power allocation in heterogeneous cloud radio access networks," *IEEE Trans. Veh. Technol.*, vol. 64, no. 11, pp. 5275-5287, Nov. 2015.