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

Tian Dang and Mugen Peng
Key Laboratory of Universal Wireless Communication, Ministry of Education
Beijing University of Posts and Telecommunications, Beijing, 100876, China
Email: {tiandang, pmg}@bupt.edu.cn

***Abstract*—Fog radio access networks (F-RANs) take full advantages of both fog and cloud computing technologies, and benefit from using network slicing to support diverse services with different quality-of-service (QoS) requirements. In this paper, to make the network slicing in F-RANs work efficiently, the radio resource allocation for different network slices is exploited in a downlink F-RAN with device-to-device communication, which is logically partitioned into a high-transmission-rate slice and a low-latency slice. A multi-objective optimization problem is presented with respect to diverse QoS demands, which can be formulated as a single-objective optimization problem with a linear weighted sum method. Then an equivalent drift-plus-penalty minimization problem with Lyapunov optimization is proposed, in which two subproblems are partitioned and solved by weighted minimum mean square error approach and Lagrange dual decomposition method, respectively. Numerical results confirm that a $[\mathcal{O}(1/V), \mathcal{O}(V)]$ utility-delay tradeoff is obtained, in which either a 34.5% reduction in the queuing delay or a 6.5% increase in the average weighted utility can be achieved in a specific model by choosing proper values of V.**

I. Introduction

As predicted by Cisco, monthly global mobile traffic will be 49 exabytes by 2021 [1]. To meet the explosive growth, the fifth generation (5G) wireless networks are presented, in which the usage scenarios are mainly classified into enhanced mobile broadband (eMBB), ultra-reliable and low-latency communication (uRLLC) and massive machine type communication. A-mong these scenarios, supporting high-transmission-rate communications for eMBB [2] and low-latency communications for uRLLC [3] has drown significant attention.

To support various quality-of-service (QoS) requirements and eliminate redundant traffic, edge computing and caching techniques [4] have been presented. In [5], a logic fog layer was introduced jointly taking the fog computing and caching abilities into consideration. In [6], a hierarchical content caching paradigm was proposed with ergodic rate and transmit latency taken into consideration, where the latency can be significantly decreased. A framework that jointly considered networking, caching, and computing techniques was proposed in [7] to meet the requirements of next generation green wireless networks. Furthermore, fog radio access networks (F-RANs) have been presented, which take full advantages of edge fog computing, centralized cloud computing and caching techniques [8].

Recently, network slicing is a proposed technology considering various QoS demands in F-RANs. On a shared infrastructure, network slicing is used to partition a physical network into several logical slices [9]. In [10], joint sub-carrier allocation and caching placement are investigated for two network slices in F-RANs with a two-step iterative algorithm. In [11], a distributed solution framework was investigated to maximize the profit while guarantee users' QoS by the joint slice and transmit power allocation. Moreover, to keep the network stable, Lyapunov optimization [12] was proposed to achieve the tradeoff between the queuing delay and the objective network performance with lower complexity.

Both high-transmission-rate slices and low-latency slices, which are referred to as eMBB and uRLLC in 5G, respectively, are concerned in F-RANs. With the consideration of delay performance, the optimization of joint power consumption, subcarrier and beamforming is studied. The contributions of this paper are as follows.

- The radio resource allocation is formulated as a multi-objective optimization problem, aiming to both maximize the transmission rate for eMBB and minimize the queuing delay for uRLLC in F-RANs.
- A single-objective function with a linear weighted sum method is first presented. Then the Lyapunov optimization is used to reformulate an equivalent drift-plus-penalty minimization problem, which is partitioned into two irrelevant subproblems, tackled with weighted minimum mean square error (WMMSE) approach and Lagrange dual decomposition method, respectively.
- Numerical results demonstrate a $[\mathcal{O}(1/V), \mathcal{O}(V)]$ utility-delay tradeoff. In a specific simulation model, choosing a proper V can bring a 34.5% reduction in the queuing delay or a 6.5% increase in the average weighted utility.

The remainder of this paper is organized as follows. In Section II, the concerned scenario is described. In Section III, the optimization model is formulated and transformed with the Lyapunov optimization, which is partitioned into two subproblems and solved with the WMMSE approach and the Lagrange dual decomposition method, respectively in Section IV. Numerical results are presented in Section V, followed by the conclusion in Section VI

978-1-5386-4328-0/18/$31.00 ©2018 IEEE

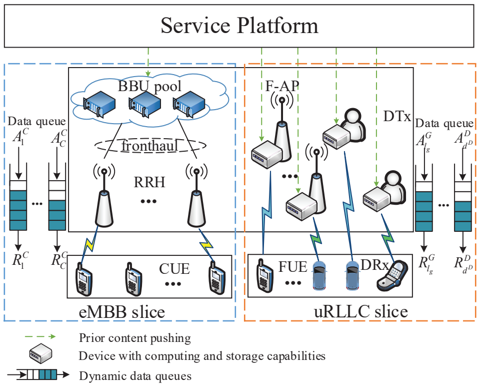

Fig. 1. A specific logical architecture of F-RANs consisting of an eMBB slice and a uRLLC slice.

For notation, calligraphy upper letters are used to denote sets. Vectors and matrices are represented by boldface lower case letters and boldface upper case letters, respectively. The conjugate transpose of $\mathbf{x}$ is denoted by $\mathbf{x}^H$. $\mathbf{I}_N$ and $\mathbf{0}_N$ represent identity matrix and zero matrix, the dimension of which is $N \times N$. $[x]^+ \overset{\Delta}{=} \max\{0, x\}$. Let $\mathbb{C}^{A\times B}$ denote the set of all $A \times B$ matrices with complex entries. The cardinality of the set $\mathcal{L}$ is donated by $|\mathcal{L}|$, whereas $|x|$ represents the absolute value of the complex scalar x. Finally, $\mathcal{CN}\left(\mu, \sigma^2\right)$ denotes the Gaussian distribution with mean μ and variance σ^2.

II. System Model

As shown in Fig. 1, the concerned F-RAN comprises an eMBB slice and a uRLLC slice. Suppose that the caching and computing scheme has been perfectly organized in fog access points (F-APs) and D2D transmitters (DTxs) to support user's requirements by local processing and transmission. In the eMMB slice, cooperative remote radio heads (RRHs) connect to centralized baseband unit (BBU) pool through fronthaul links. The queue state information (QSI) at each user equipment (UE) is represented by the data queue. Suppose the total bandwidth W is orthogonally divided into ηW and $(1-\eta)W$, $\eta \in [0, 1]$ for eMBB and uRLLC slices, respectively. The time dimension is partitioned into slots indexed by $t \in \{0, 1, 2, ...\}$ and the duration of each slot is τ.

A. System Model for eMBB Slice

Suppose there are C single-antenna cloud-UEs (CUEs) and M RRHs and each RRH equips with N antennas. Denote by $\mathcal{C} = \{1, 2, \ldots, C\}$ and $\mathcal{M} = \{1, 2, \ldots, M\}$ the set of CUEs and RRHs. The signal received at CUE c in time slot t is

$$\begin{aligned} y_c^C(t) = \mathbf{h}_c^H(t)\,\mathbf{w}_c(t)\,s_c^C(t) \\ + \sum_{i\in\mathcal{C}, i\neq c} \mathbf{h}_c^H(t)\,\mathbf{w}_i(t)\,s_i^C(t) + z_c^C(t), \end{aligned} \quad (1)$$

where $\mathbf{h}_c(t) \in \mathbb{C}^{MN\times 1}$ is the channel state information (CSI) vector from RRHs to CUE c, $\mathbf{w}_c(t) \in \mathbb{C}^{MN\times 1}$ denotes the beamforming vector for CUE c, $s_c^C(t)$ represents the information symbol sent to CUE c with $\mathbb{E}[|s_c^C(t)|^2] = 1$, and $z_c^C(t) \sim \mathcal{CN}\left(0, \sigma^2\right)$ is the additive white Gaussian noise. The downlink transmission rate of CUE c can be expressed as

$$R_c^C(t) = \eta W \log_2\Bigg(1 + \frac{\left|\mathbf{h}_c^H(t)\,\mathbf{w}_c(t)\right|^2}{\sum\limits_{i\in\mathcal{C}, i\neq c} \left|\mathbf{h}_c^H(t)\,\mathbf{w}_i(t)\right|^2 + \sigma^2}\Bigg). \quad (2)$$

The transmit power consumption per RRH is written as $P_m^R(t) = \sum_{c\in\mathcal{C}} \mathbf{w}_c^H(t)\,\mathbf{D}_m^H \mathbf{D}_m \mathbf{w}_c(t)$, where $\mathbf{D}_m = (\underbrace{\mathbf{0}_N, \ldots, \mathbf{0}_N}_{m-1}, \mathbf{I}_N, \mathbf{0}_N, \ldots, \mathbf{0}_N)$.

Let $A_c^C(t) \in [0, A_{\max}^C]$ denote the random arrival rate (Mbit/slot) at CUE c. Suppose $A_c^C(t)$ is independent identically distributed (i.i.d.) over all slots. Denote by $Q_c^C(t)$ the data queue of CUE c at slot t. The queue dynamic at CUE c is expressed as $Q_c^C(t+1) = \left[Q_c^C(t) + A_c^C(t) - R_c^C(t)\,\tau\right]^+$.

The minimum allowable transmission rate of CUEs is denoted by R^{req}, and $R_c^C(t)$ must satisfy

$$R_c^C(t) \geq R^{req}. \quad (3)$$

Denote by $P^{R,\max}$ the maximum allowable transmit power consumption. For each RRH, $P_m^R(t)$ is constrained by

$$P_m^R(t) \leq P^{R,\max}. \quad (4)$$

Due to the following definition, the queue stability of all CUEs can be expressed as

$$\lim_{t\to\infty} \frac{\mathbb{E}\{|Q_c^C(t)|\}}{t} = 0, \quad (5)$$

Definition 1: A discrete time process $Q(t)$ is mean rate stable [12] if

$$\lim_{t\to\infty} \frac{\mathbb{E}\{|Q(t)|\}}{t} = 0. \quad (6)$$

Since the key performance indicator is spectral efficiency (SE), the objective function at slot t is calculated as $f_e(t) = \sum_{c\in\mathcal{C}} R_c^C(t)\,\tau$, which represents the total amount of data transmitted to CUEs in time slot t. For convenience, $f_e(t)$ is defined as the utility of the eMBB slice at time slot t.

B. System Model for uRLLC Slice

Suppose there are G F-APs, D DTxs with the set denoted by $\mathcal{G} = \{1, 2, \ldots. G\}$ and $\mathcal{D} = \{1, 2, \ldots, D\}$, respectively. Denote by $l_g \in \mathcal{L}_g$ the fog-UE (FUE) served by F-AP g and $\mathcal{L}_g$ represents the set. Denote by d^D the D2D receiver (DRx) served by DTx d. The spectrum resource is orthogonally divided into S subcarriers equally, with the bandwidth of each subcarrier as $W_1 \overset{\Delta}{=} (1-\eta)\,W/S$. The set of subcarriers is denoted by $\mathcal{S} = \{1, 2, \ldots, S\}$.

For each FUE $l_g \in \mathcal{L}_g$, the received signal is given by

$$\begin{aligned} y_{l_g}^G(t) = \sum_s a_{s,l_g}^G(t)\sqrt{P_{s,l_g}^G(t)\,h_{s,l_g}^G(t)}\,s_{l_g}^G(t) + z_{l_g}^G(t) \\ + \sum_{s,d} b_{s,d^D}^D(t)\sqrt{P_{s,d^D}^D(t)\,h_{d,s,l_g}^{DG}(t)}\,s_{d_D}^D(t), \end{aligned} \quad (7)$$

where $P_{s,l_g}^G(t)$ and $h_{s,l_g}^G(t)$ are the transmit power consumption and the channel fading power gain from F-AP g to FUE

l_g on subcarrier s, $a_{s,l_g}^G(t)$ and $b_{s,d^D}^D(t) \in \{0,1\}$ denote the subcarrier allocation indicators, $P_{s,d^D}^D(t)$ and $h_{d,s,l_g}^{DG}(t)$ represent the transmit power consumption to DRx d^D and the channel fading power gain from DTx d to FUE l_g on subcarrier s, $s_{l_g}^G(t)$ and $s_{d^D}^D(t)$ denote the information symbols sent to FUE l_g and DRx d^D with $\mathbb{E}[|s_{l_g}^G(t)|^2] = 1$ and $\mathbb{E}[|s_{d^D}^D(t)|^2] = 1$, and $z_{l_g}^G(t) \sim \mathcal{CN}\left(0,\sigma^2\right)$ is the additive white Gaussian noise at FUE l_g. The signal-to-interference-plus-noise ratio (SINR) of FUE l_g can be expressed as $SINR_{l_g}^G(t) = \frac{a_{s,l_g}^G(t)P_{s,l_g}^G(t)h_{s,l_g}^G(t)}{\sum_d b_{s,d^D}^D(t)P_{s,d^D}^D(t)h_{d,s,l_g}^{DG}(t)+\sigma^2}$. Then $R_{l_g}^G(t) = W_1\sum_{s\in\mathcal{S}}\log_2(1+SINR_{l_g}^G(t))$ denotes the downlink transmission rate to FUE l_g. The transmit power per F-AP is $P_g^G(t) = \sum_{s\in\mathcal{S},l_g\in\mathcal{L}_g} a_{s,l_g}^G(t)\,P_{s,l_g}^G(t)$.

For each DRx d^D, the received signal is given by

$$y_{d^D}^D(t) = \sum_s b_{s,d^D}^D(t)\sqrt{P_{s,d^D}^D(t)\,h_{s,d^D}^D(t)}s_{d^D}^D(t) + z_{d^D}^D(t) + \sum_{s,g,l_g} a_{s,l_g}^G(t)\sqrt{P_{s,l_g}^G(t)\,h_{g,s,d^D}^{GD}(t)}s_{l_g}^G(t), \quad (8)$$

where $h_{s,d^D}^D(t)$ and $h_{g,s,d^D}^{GD}(t)$ are the channel fading power gain from DTx d and F-AP g to DRx d^D on subcarrier s, and $z_{d^D}^D(t) \sim \mathcal{CN}\left(0,\sigma^2\right)$ is the additive white Gaussian noise at DRx d^D. The SINR of DRx d^D can be expressed as $SINR_{d^D}^D(t) = \frac{b_{s,d^D}^D(t)P_{s,d^D}^D(t)h_{s,d^D}^D(t)}{\sum_{g,l_g} a_{s,l_g}^G(t)P_{s,l_g}^G(t)h_{g,s,d^D}^{GD}(t)+\sigma^2}$. The downlink transmission rate of DRx d^D is $R_{d^D}^D(t) = W_1\sum_{s\in\mathcal{S}}\log_2(1+SINR_{d^D}^D(t))$. The transmit power consumption per DTx is $P_d^D(t) = \sum_{s\in\mathcal{S}} b_{s,d^D}^D(t)\,P_{s,d^D}^D(t)$.

Let $A_{l_g}^G(t) \in [0, A_{\max}^G]$ and $A_{d^D}^D(t) \in [0, A_{\max}^D]$ denote the random arrival rates (Mbit/slot) at FUE l_g and DRx d^D, respectively. Suppose $A_{l_g}^G(t)$ and $A_{d^D}^D(t)$ are i.i.d. over all slots. Denote by $Q_{l_g}^G(t)$ and $Q_{d^D}^D(t)$ the data queues of FUE l_g and DRx d^D at time slot t. The queue dynamics are expressed as $Q_{l_g}^G(t+1) = [Q_{l_g}^G(t) + A_{l_g}^G(t) - R_{l_g}^G(t)\tau]^+$ and $Q_{d^D}^D(t+1) = [Q_{d^D}^D(t) + A_{d^D}^D(t) - R_{d^D}^D(t)\tau]^+$.

Denoted by $P^{G,\max}$ and $P^{D,\max}$ the maximum allowable transmit power consumption of F-APs and DTxs, and $P_g^G(t)$ and $P_d^D(t)$ are constrained by

$$P_g^G(t) \le P^{G,\max}, \quad (9)$$

$$P_d^D(t) \le P^{D,\max}. \quad (10)$$

To meet the QoS demand, probabilistic constraints on the queue length with a probability $\varepsilon \in (0,1)$ and an allowable upper bound Q^{req} [13] are given by

$$\lim_{t\to\infty} \Pr\left[Q_{l_g}^G(t) \ge Q^{req}\right] \le \varepsilon, \quad (11)$$

$$\lim_{t\to\infty} \Pr\left[Q_{d^D}^D(t) \ge Q^{req}\right] \le \varepsilon. \quad (12)$$

The objective function can be calculated as $f_u(t) = \sum_{g\in\mathcal{G},l_g\in\mathcal{L}_g} Q_{l_g}^G(t) + \sum_{d\in\mathcal{D}} Q_{d^D}^D(t)$, which denotes the total queue length in time slot t. For convenience, $f_u(t)$ is defined as the utility of the uRLLC slice at time slot t.

III. Problem Formulation

The average utilities are $\overline{f}_e = \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[f_e(t)\right]$ and $\overline{f}_u = \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[f_u(t)\right]$. A radio resource allocation optimization problem can be formulated as

$$\begin{aligned} \text{(P0)} \quad & \max_{\{\mathbf{w}_c,\boldsymbol{\pi},\mathbf{P}\}} : \left[\overline{f}_e, -\overline{f}_u\right] \\ \text{s.t.} \quad & (3)-(5),(9)-(12), \\ & \sum_{g,l_g} a_{s,l_g}^G(t) = 1, a_{s,l_g}^G(t) \in \{0,1\}, \quad (13) \\ & \sum_d b_{s,d^D}^D(t) = 1, b_{s,d^D}^D(t) \in \{0,1\}, \quad (14) \end{aligned}$$

where $\boldsymbol{\pi}$ denotes the combination vector of $a_{s,l_g}^G(t)$ and $b_{s,d^D}^D(t)$, $\mathbf{P}$ is the combination vector of $P_{s,l_g}^G(t)$ and $P_{s,d^D}^D(t)$. (13) and (14) denote the subcarrier allocation limitations.

A. Objective and Constraint Transformation

Problem (P0) can be reformulated into a single-objective optimization problem with a linear weighted sum method, where the single-objective function is the average weighted utility $\overline{u} = \kappa\overline{f}_e - \overline{f}_u = \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\left(\kappa f_e(t) - f_u(t)\right)$. $\kappa > 0$ is the weight parameter representing the importance of $f_e(t)$. Define $u(t) \triangleq \kappa f_e(t) - f_u(t)$ as the weighted utility.

Assume each term of $\mathbf{h}_c^H(t)\,\mathbf{w}_c(t)$ has zero imaginary part, which is achieved by rotating the phase of $\mathbf{w}_c(t)$ [14]. Thus (3) can be rewritten as a second-order cone (SOC) constraint:

$$\frac{\left|\mathbf{h}_c^H(t)\,\mathbf{w}_c(t)\right|^2}{2^{R^{req}/\eta W} - 1} \ge \Big(\sum_{i\neq c, i\in\mathcal{C}}\left|\mathbf{h}_c^H(t)\,\mathbf{w}_i(t)\right|^2 + \sigma^2\Big). \quad (15)$$

Using Markov's inequality [15], (11) and (12) can be linearized such that $\overline{Q}_{l_g}^G \le Q^{req}\varepsilon$ and $\overline{Q}_{d^D}^D \le Q^{req}\varepsilon$. Denote by $H_{l_g}(t)$ and $F_{d^D}(t)$ the virtual queues, evolving as $H_{l_g}(t+1) = [H_{l_g}(t) + Q_{l_g}^G(t) - Q^{req}\varepsilon]^+$ and $F_{d^D}(t+1) = \left[F_{d^D}(t) + Q_{d^D}^D(t) - Q^{req}\varepsilon\right]^+$. Both (11) and (12) are satisfied if $H_{l_g}(t)$ and $F_{d^D}(t)$ are mean rate stable:

$$\lim_{t\to\infty}\frac{\mathbb{E}\{|H_{l_g}(t)|\}}{t} = 0, \quad (16)$$

$$\lim_{t\to\infty}\frac{\mathbb{E}\{|F_{d^D}(t)|\}}{t} = 0. \quad (17)$$

Then, problem (P0) can be rewritten as

$$\begin{aligned} \text{(P1)} \quad & \max_{\{\mathbf{w}_c,\boldsymbol{\pi},\mathbf{P}\}} : \overline{u} \\ \text{s.t.} \quad & (4),(5),(9),(10),(13)-(17). \end{aligned}$$

B. Lyapunov Optimization

Lyapunov optimization is utilized to transform (P1) into a per-slot drift-plus-penalty minimization problem. Let $\boldsymbol{\Theta}(t) \triangleq \left(\mathbf{Q}^C(t), \mathbf{H}(t), \mathbf{F}(t)\right)$ denote the vector of all queues, where $\mathbf{Q}^C(t)$, $\mathbf{H}(t)$ and $\mathbf{F}(t)$ are the vectors of $Q_c^C(t)$, $H_{l_g}(t)$ and $F_{d^D}(t)$, respectively. Lyapunov function [12], which denotes the scalar measure of network queue congestion, is given by

$$L\left(\boldsymbol{\Theta}(t)\right) \triangleq \frac{1}{2}\Big(\sum_c Q_c^C(t)^2 + \sum_{g,l_g} H_{l_g}^2(t) + \sum_d F_{d^D}^2(t)\Big).$$

Denote by $\Delta\left(\mathbf{\Theta}\left(t\right)\right)$ the Lyapunov drift:

$$\Delta\left(\mathbf{\Theta}\left(t\right)\right) \triangleq \mathbb{E}\left\{L\left(\mathbf{\Theta}\left(t+1\right)\right) - L\left(\mathbf{\Theta}\left(t\right)\right)\left|\mathbf{\Theta}\left(t\right)\right.\right\}, \qquad (18)$$

which represents the change in $L\left(\mathbf{\Theta}\left(t\right)\right)$ from one slot to the next and is utilized to keep queues stable associated with (P1).

Lemma 1: For any slot t, with observed CSI and arrival rates, the Lyapunov drift (18) satisfies:

$$\begin{aligned}
&\Delta\left(\mathbf{\Theta}\left(t\right)\right) \qquad (19)\\
&\leq B - \sum_{c}(Q_c^C\left(t\right) + A_c^C\left(t\right))\mathbb{E}\{R_c^C\left(t\right)\tau\left|\mathbf{\Theta}\left(t\right)\right.\}\\
&- \sum_{g,l_g}(Q_{l_g}^G\left(t\right) + A_{l_g}^G\left(t\right) + \frac{1}{2}H_{l_g}\left(t\right))\mathbb{E}\{R_{l_g}^G\left(t\right)\tau\left|\mathbf{\Theta}\left(t\right)\right.\}\\
&- \sum_{d}(Q_{d^D}^D\left(t\right) + A_{d^D}^D\left(t\right) + \frac{1}{2}F_{d^D}\left(t\right))\mathbb{E}\{R_{d^D}^D\left(t\right)\tau\left|\mathbf{\Theta}\left(t\right)\right.\},
\end{aligned}$$

where $B > 0$ is a finite constant.

Instead of optimizing the average weighted utility while minimizing the upper bound of $\Delta\left(\mathbf{\Theta}\left(t\right)\right)$, we minimize a bound on a drift-plus-penalty expression $\Delta\left(\mathbf{\Theta}\left(t\right)\right) - V\mathbb{E}\left\{u\left(t\right)\right|\mathbf{\Theta}\left(t\right)\}$ at each time slot t, where $V > 0$ is an emphasis weight. Moreover, according to [12], we have

$$\Delta\left(\mathbf{\Theta}\left(t\right)\right) - V\mathbb{E}\{u(t)|\mathbf{\Theta}(t)\} \leq B - Vu^{opt} - \sum_{c}\epsilon Q_c^C\left(t\right),$$

where B, ϵ and V are positive constants for all slots and u^{opt} is the theoretical optimal value of $\overline{u}$. Then the average weighted utility and the average queue length satisfy

$$\overline{u} \geq u^{opt} - \frac{B}{V}, \quad \overline{Q}_c^C \leq \frac{B + V(u^{\max} - u^{opt})}{\epsilon} \qquad (20)$$

Remark 1: A $[\mathcal{O}(1/V), \mathcal{O}(V)]$ tradeoff between the average weighted utility and the average queue backlog is achieved.

For simplicity, some coefficients are defined as

$$\phi_c^C\left(t\right) \triangleq (Q_c^C\left(t\right) + A_c^C\left(t\right) + V\kappa)\tau, \qquad (21)$$

$$\alpha_{l_g}^G\left(t\right) \triangleq (Q_{l_g}^G\left(t\right) + A_{l_g}^G\left(t\right) + \frac{1}{2}H_{l_g}\left(t\right))\tau, \qquad (22)$$

$$\beta_{d^D}^D\left(t\right) \triangleq (Q_{d^D}^D\left(t\right) + A_{d^D}^D\left(t\right) + \frac{1}{2}F_{d^D}\left(t\right))\tau. \qquad (23)$$

The drift-plus-penalty optimization problem is given as

$$\begin{aligned}
\text{(P2)} \quad \max_{\{\mathbf{w}_c, \boldsymbol{\pi}, \mathbf{P}\}} : &\sum_{c}\phi_c^C\left(t\right)R_c^C\left(t\right) + \sum_{g,l_g}\alpha_{l_g}^G\left(t\right)R_{l_g}^G\left(t\right)\\
&+ \sum_{d\in\mathcal{D}}\beta_{d^D}^D\left(t\right)R_{d^D}^D\left(t\right)\\
\text{s.t.} \quad &(4), (9), (10), (13) - (15).
\end{aligned}$$

IV. Radio Resource Allocation Optimization

Problem (P2) can be partitioned into two subproblems:

$$\begin{aligned}
\text{(P2-1)} \quad &\max_{\{\mathbf{w}_c\}} : \sum_{c}\phi_c^C\left(t\right)R_c^C\left(t\right)\\
&\text{s.t.} \quad (4), (15).\\
\text{(P2-2)} \quad &\max_{\{\boldsymbol{\pi}, \mathbf{P}\}} : \sum_{g,l_g}\alpha_{l_g}^G\left(t\right)R_{l_g}^G\left(t\right) + \sum_{d}\beta_{d^D}^D\left(t\right)R_{d^D}^D\left(t\right)\\
&\text{s.t.} \quad (9), (10), (13), (14).
\end{aligned}$$

A. Beamforming Design

The weighted sum-rate (WSR) maximization problem in (P2-1) can be reformulated as a WMMSE problem and a local optimum is achieved with the block coordinate descent method. The equivalence between the WSR maximization problem and the WMMSE problem is stated as follows.

Proposition 1: Problem (P2-1) has the same optimal solution as the following WMMSE minimization problem:

$$\begin{aligned}
\min_{\{\mathbf{w}_c, v_c, u_c\}} : &\sum_{c}\phi_c^C\left(t\right)\left(v_c\left(t\right)e_c\left(t\right) - \log v_c\left(t\right)\right) \qquad (24)\\
\text{s.t.} \quad &(4), (15),
\end{aligned}$$

where $v_c(t)$ is the mean-square error (MSE) weight for CUE c at slot t, and $e_c\left(t\right) \triangleq \mathbb{E}\{\left(u_k\left(t\right)y_c^C\left(t\right) - s_c^C\left(t\right)\right)^2\}$ is the corresponding MSE under the receiver $u_c(t) \in \mathbb{C}$.

This optimization problem (24) can be settled through the block coordinate descent method by iterating among $\mathbf{w}_c(t)$, $v_c(t)$, and $u_c(t)$. For simplicity, the symbol t is omitted in the following iteration.

- The optimal MSE weight v_c under fixed $\mathbf{w}_c$ and u_c is given by $v_c = e_c^{-1}$.
- The optimal receiver u_c under fixed $\mathbf{w}_c$ and v_c is given by $u_c = \mathbf{h}_c^H\mathbf{w}_c\{\sum_{i\in\mathcal{C}}\mathbf{w}_i^H\mathbf{h}_c\mathbf{h}_c^H\mathbf{w}_i + \sigma^2\}^{-1}$.
- The beamforming vector design solution can be observed by solving a SOC programming optimization problem under fixed v_c and u_c:

$$\begin{aligned}
\min_{\{\mathbf{w}_c\}} : &\sum_{i\in\mathcal{C}}\mathbf{w}_i^H\{\sum_{c\in\mathcal{C}}\phi_c^C v_c u_c^H\mathbf{h}_c\mathbf{h}_c^H u_c\}\mathbf{w}_i \qquad (25)\\
&- 2\sum_{c\in\mathcal{C}}\phi_c^C v_c \mathrm{Re}\left\{u_c\mathbf{h}_c^H\mathbf{w}_c\right\}\\
\text{s.t.} \quad &(4), (15).
\end{aligned}$$

Problem (25) can be solved by using the Matlab software for disciplined convex programming.

B. Power and Subcarrier Allocation

For simplicity, the symbol t is omitted during solving problem (P2-2) at slot t. Denote by $I_{DG}^{\max}$ and $I_{GD}^{\max}$ the maximum allowable interference on FUEs and DRxs, respectively. The the transmission rates on FUEs and DRxs with $I_{DG}^{\max}$ and $I_{GD}^{\max}$ replacing the interference are represented by $\tilde{R}_{l_g}^G$ and $\tilde{R}_{d^D}^D$, which satisfy $\tilde{R}_{l_g}^G \leq R_{l_g}^G$ and $\tilde{R}_{d^D}^D \leq R_{d^D}^D$.

Then, problem (P2-2) can be approximatively rewritten as a convex problem:

$$\begin{aligned}
\max_{\{\boldsymbol{\pi}, \mathbf{P}\}} : &\sum_{g,l_g}\alpha_{l_g}^G\tilde{R}_{l_g}^G + \sum_{d\in\mathcal{D}}\beta_{d^D}^D\tilde{R}_{d^D}^D \qquad (26)\\
\text{s.t.} \quad &(9), (10), (13), (14),\\
&\sum_{d}b_{s,d^D}^D P_{s,d^D}^D h_{d,s,l_g}^{DG} \leq I_{DG}^{\max}, \qquad (27)\\
&\sum_{g,l_g}a_{s,l_g}^G P_{s,l_g}^G h_{g,s,d^D}^{GD} \leq I_{GD}^{\max}. \qquad (28)
\end{aligned}$$

With each constraint in problem (26) linear in terms of the transmit power consumption, strong duality holds between

(26) and its dual problem which be solved with the Lagrange dual decomposition method. The computational procedure is omitted for convenience and similar procedure can be found in [16]. The optimal power allocation in each iteration is given by

$$P_{s,l_g}^{G} = \Big[\frac{\alpha_{l_g}^{G}}{(\gamma_g + \sum\limits_{d} \varphi_{s,d^D} h_{g,s,d^D}^{GD}) \ln 2} - \frac{I_{DG}^{\max} + \sigma^2}{h_{s,l_g}^{G}}\Big]^{+}, \quad (29)$$

$$P_{s,d^D}^{D} = \Big[\frac{\beta_{d^D}^{D}}{(\mu_d + \sum\limits_{g,l_g} \chi_{s,l_g} h_{d,s,l_g}^{DG}) \ln 2} - \frac{I_{GD}^{\max} + \sigma^2}{h_{s,d^D}^{D}}\Big]^{+}, \quad (30)$$

where γ_g, μ_d, χ_{s,l_g} and φ_{s,d^D} are the Lagrangian multipliers of this iteration associated with (9), (10), (27) and (28). For simplicity, we have

$$\Omega_{s,l_g}^{G} = (\gamma_g + \sum_{d} \varphi_{s,d^D} h_{g,s,d^D}^{GD})(I_{DG}^{\max} + \sigma^2)/h_{s,l_g}^{G}, \quad (31)$$

$$\Omega_{s,d^D}^{D} = (\mu_d + \sum_{g,l_g} \chi_{s,l_g} h_{d,s,l_g}^{DG})(I_{GD}^{\max} + \sigma^2)/h_{s,d^D}^{D}. \quad (32)$$

The optimal subcarrier allocation solution can be given by

$$a_{s,l_g}^{G} = \begin{cases} 1, & (g,l_g) = \arg\max\limits_{g,l_g} f_{s,l_g}^{G}, \\ 0, & \text{otherwise}, \end{cases} \quad (33)$$

$$b_{s,d^D}^{D} = \begin{cases} 1, & d = \arg\max\limits_{d} f_{s,d^D}^{D}, \\ 0, & \text{otherwise}, \end{cases} \quad (34)$$

where f_{s,l_g}^{G} and f_{s,d^D}^{D} are defined as

$$f_{s,l_g}^{G} = \alpha_{l_g}^{G}\Big[\log_2(\frac{\alpha_{l_g}^{G}}{\Omega_{s,l_g}^{G} \ln 2})\Big]^{+} - \Big[\frac{\alpha_{l_g}^{G}(t)}{\ln 2} - \Omega_{s,l_g}^{G}\Big]^{+}, \quad (35)$$

$$f_{s,d^D}^{D} = \beta_{d^D}^{D}\Big[\log_2(\frac{\beta_{d^D}^{D}}{\Omega_{s,d^D}^{D} \ln 2})\Big]^{+} - \Big[\frac{\beta_{d^D}^{D}}{\ln 2} - \Omega_{s,d^D}^{D}\Big]^{+}. \quad (36)$$

V. Numerical Results

In this section, the tradeoff between the weighted utility and the queue backlog for network slicing in F-RANs is evaluated. There are $C = 10$ CUEs and $M = 3$ RRHs in the eMBB slice, where $R^{req} = 1.5$ Mbit/s and $P^{R,\max} = 27$ dBm. Each RRH equips with $N = 2$ antennas. There are $G = 3$ F-APs and $D = 6$ DTxs in the uRLLC slice, where 6 DRxs are paired up with DTxs, each F-AP g serves $|\mathcal{L}_g| = 2$ FUEs, $P^{G,\max} = 27$ dBm and $P^{D,\max} = 25$ dBm. All devices are located on a circle with radius 100 meters, and occupy $W = 20$ Mega Hertz (MHz) bandwidth resource. For convenience, the arrival rate of each UE is supposed to be constant and represented as λ_1, λ_2 and λ_3 for CUEs, FUEs and DRxs, respectively. Specifically, we fix $\tau = 0.15$ seconds and $\kappa = 1$.

A. Queue Stability Evaluation

The queue stability performance of CUEs is evaluated in this section. The value of the traffic data arrival rate is set to $\lambda_1 \in \{0.6, 0.63, 0.66\}$ Mbit/slot at each time slot and η is set to 0.5. The average queue length versus t under different V and λ_1 is shown in Fig. 2. It can be observed that the average queue length becomes stable over a period of time and then fluctuates around a certain value with t. Under the same V, a larger arrival rate leads to a larger average queue length, since different arrival rates cause different transmit power consumptions which are limited on each RRH. A larger V leads to a larger average queue length. This happens because a larger V makes us put more emphasis on the weighted utility, resulting in a larger stable value of the queue length. Moreover, it can be observed that under a specific arrival rate, i.e. $\lambda_1 = 0.66$ Mbit/slot, choosing a proper V can bring a 34.5% reduction in the queuing delay.

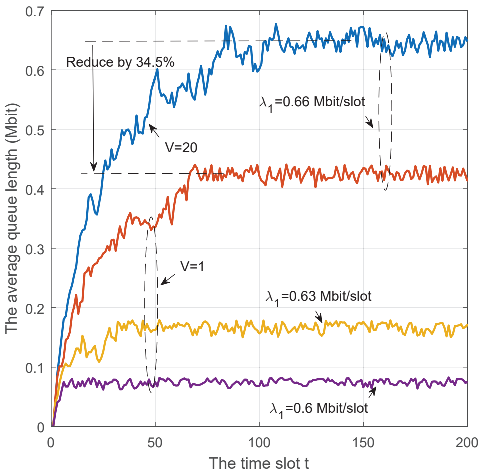

Fig. 2. The average queue length versus the simulation time.

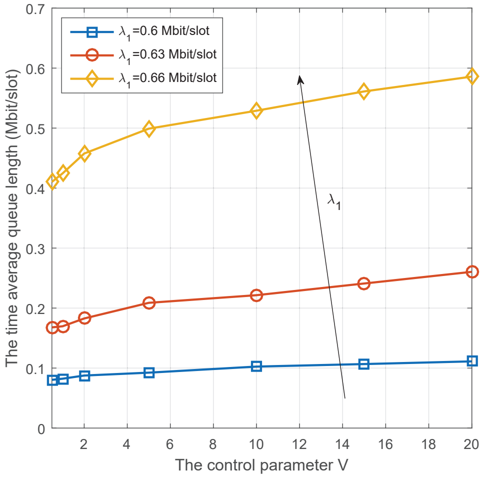

Fig. 3. The time average queue length versus the control parameter.

B. The Utility-Delay Tradeoff

The average queue backlog versus V under different arrival rates $\lambda_1 \in \{0.6, 0.63, 0.66\}$ Mbit/slot is shown in Fig. 3, where the average queue backlog linearly increases with V for any arrival rate. It can be understood by the fact that greater emphasis is put on the weighted utility with increasing V, resulting in a larger queue length.

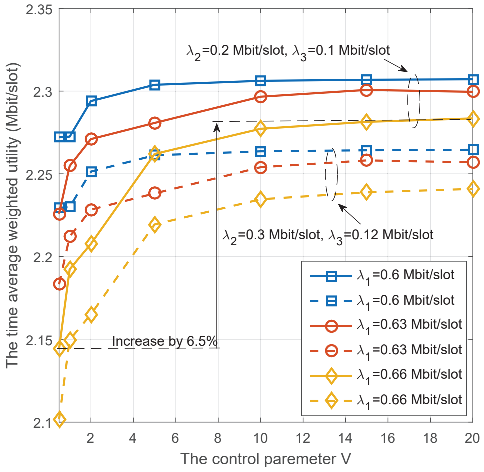

Fig. 4. The time average weighted utility versus the control parameter.

The average weighted utility versus V under different arrival rates is shown in Fig. 4. It can be observed that the average weighted utility increases with V for any arrival rate. This happens because greater emphasis is put on the weighted utility when V increases. Moreover, under specific arrival rates, i.e. $(\lambda_1, \lambda_2, \lambda_3) = (0.66, 0.2, 0.1)$ Mbit/slot, choosing a proper V brings a 6.5% increase in the average weighted utility. Both higher arrival rates of CUEs and lower arrival rates of UEs in the uRLLC slice will lead to a larger average weighted utility. It can be intuitively understood based on the definition of the average weighted utility, in which a larger arrival rate leads to a higher sum transmission rate in the eMBB slice and a larger sum queue backlog in the uRLLC slice. Fig. 4 combined with Fig. 3 illustrates the utility-delay tradeoff proposed in *Remark* 1.

VI. Conclusion

In this paper, a delay-aware radio resource allocation for network slicing has been concerned in a fog radio access network. A multi-objective optimization problem has been formulated to both maximize the average sum rate for a high-transmission-rate slice and minimize the average total queuing delay for a low-latency slice. The multi-objective function has been transformed into a weighted utility function and reformulated as a drift-plus-penalty minimization problem with Lyapunov optimization. Then this problem has been partitioned into two subproblems, solved with the weighted minimum mean square error approach and the Lagrange dual decomposition method, respectively. A $[\mathcal{O}(1/V), \mathcal{O}(V)]$ utility-delay tradeoff has been achieved and the simulation results have shown that in a specific simulation model, choosing a proper V can bring either a 34.5% reduction in the queuing delay or a 6.5% increase in the average weighted utility.

Acknowledgment

This work was supported in part by the State Major Science and Technology Special Project under Grant 2016ZX03001020-006 and 2017ZX03001025-006, and in part by the National Program for Special Support of Eminent Professionals.

References

[1] Cisco System, “Cisco visual networking index: Global mobile data traffic forecast update, 2016-2021,” Cisco System, White Paper, Feb. 2017.

[2] K. Zhang, M. Peng, P. Zhang, and X. Li, “Secrecy-optimized resource allocation for device-to-device communication underlaying heterogeneous networks,” *IEEE Trans. Veh. Technol.*, vol. 66, no. 2, pp. 1822-1834, Feb. 2017.

[3] G. M. S. Rahman, M. Peng, and K. Zhang, “Radio resource allocation for achieving ultra-low latency in fog radio access networks,” *IEEE Access*, DOI: 10.1109/ACCESS.2018.2805303, 2018.

[4] C. Fang, F. R. Yu, T. Huang, J. Liu, and Y. Liu, “A survey of energy-efficient caching in information-centric networking,” *IEEE Commun. Mag.*, vol. 52, no. 11, pp. 122-129, Nov. 2014.

[5] M. Peng, S. Yan, K. Zhang, and C. Wang, “Fog-computing-based radio access networks: Issues and challenges,” *IEEE Network*, vol. 30, no. 4, pp. 46-53, Jul. 2016.

[6] S. Jia, Y. Ai, Z. Zhao, M. Peng, and C. Hu, “Hierarchical content caching in fog radio access networks: Ergodic rate and transmit latency,” *China Commun.*, vol. 13, no. 12, pp. 1-14, Dec. 2016.

[7] R. Huo, F. R. Yu, T. Huang, R. Xie, J. Liu, V. C. M. Leung, and Y. Liu, “Software defined networking, caching, and computing for green wireless networks,” *IEEE Commun. Mag.*, vol. 54, no. 11, pp. 185-193, Nov. 2016.

[8] M. Peng and K. Zhang, “Recent advances in fog radio access networks: Performance analysis and radio resource allocation,” *IEEE Access*, vol. 4, pp. 5003-5009, Aug. 2016.

[9] H. Xiang, W. Zhou, M. Daneshmand, and M. Peng, “Network slicing in fog radio access networks: Issues and challenges, ” *IEEE Commun. Mag.*, vol. 55, no. 12, pp. 110-116, Dec. 2017.

[10] L. Tang, X. Zhang, H. Xiang, Y. Sun, and M. Peng, “Joint resource allocation and caching placement for network slicing in fog radio access networks, ” *2017 IEEE 18th Int. Workshop on Signal Process. Advances in Wireless Commun. (SPAWC)*, Sapporo, Japan, Jul. 2017, pp. 1-6.

[11] T. LeAnh, N. Tran, D. Ngo, and C. Hong, “Resource allocation for virtualized wireless networks with backhaul constraints,” *IEEE Commun. Lett.*, vol. 21, no. 1, pp. 148-151, Jan. 2017.

[12] M. Neely, *Stochastic Network Optimization with Application to Communication and Queuing Systems*. San Rafael, CA, USA: Morgan & Claypool, 2010.

[13] T. Vu, C. Liu, M. Bennis, M. Debbah, M. Latvaaho, and C. Hong, “Ultra-reliable and low latency communication in mmWave-enabled massive MIMO networks,” *IEEE Commun. Lett.*, vol. 21, no. 9, pp. 2041-2044, Sep. 2017.

[14] 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.

[15] A. Mukherjee, “Queue-aware dynamic on/off switching of small cells in dense heterogeneous networks,” *2013 IEEE Globecom Workshops (GC Wkshps)*, Atlanta, GA, USA, Dec. 2013, pp. 182-187.

[16] M. Peng, K. Zhang, J. Jiang, J. Wang, and W. Wang, “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.