# Joint Mode Selection and Resource Allocation for Downlink Fog Radio Access Networks Supported D2D

Hongyu Xiang, Mugen Peng, Yuanyuan Cheng, and Hsiao-Hwa Chen
Key Laboratory of Universal Wireless Communication, Ministry of Education,
Beijing University of Posts and Telecommunications, Beijing, 100876, China
Email: xianghongyu88@gmail.com
Department of Engineering Science, National Cheng Kung University, Taiwan. E-mail: hshwchen@ieee.org

***Abstract*—Presented as an innovative paradigm incorporating the cloud computing into radio access network, cloud radio access networks (C-RANs) have been shown advantageous in curtailing the capital and operating expenditures as well as providing better services to the customers. However, heavy burden on the non-ideal fronthaul limits performances of C-RANs. Here we focus on the alleviation of burden on the fronthaul via the edge devices' caches and propose a fog computing based RAN (F-RAN) architecture with three candidate transmission modes: device to device, local distributed coordination, and global C-RAN. Followed by the proposed simple mode selection scheme, the average energy efficiency (EE) of systems optimization problem considering congestion control is presented. Under the Lyapunov framework, the problem is reformulated as a joint mode selection and resource allocation problem, which can be solved by block coordinate descent method. The mathematical analysis and simulation results validate the benefits of F-RAN and an EE-delay tradeoff can be achieved by the proposed algorithm.**

## I. Introduction

To deal with the dilemma about increasing demand for capacity and reducing capital and operating expenditures, the cloud radio access network (C-RAN) is proposed as a new mobile network structure [1], which includes centralized processing base-band units (BBU) pool, large-scale cooperative transmission via distributed remote radio heads (RRHs), and fronthaul transport network connecting the BBUs and RRHs. Intense research effort on C-RANs has established that significant performance gains, like capacity and coverage improvement via cost effective approaches, can be obtained [2]. However, confronted with demand for a more aggressive fifth generation (5G) system [3], the traditional C-RAN couldn't meet the engineering requirements with the challenges of heavy traffic burden in the constrained fronthaul.

It is noted that an important share of the traffic explosion is due to duplicate downloads of a few popular contents with large sizes [4]. Therefore, numerous investigations have been done on reducing the duplicate content transmissions through intelligent caching strategies inside the mobile networks. A simple cost model is proposed in [4] to explore the potential of forward caching in cellular networks and it is shown that a proper caching will reduce mobile traffic by one third. A similar conclusion was also obtained in long term evolution networks [5] and an overview of emerging caching techniques for current mobile networks and future 5G networks is provided [6]. However, most of literatures focused on the benefit of cache incorporation, while few works have been done on how to explore cache in mobile networks and a framework with cache incorporation is needed.

In this paper, we proposed a fog computing based RAN (F-RAN) architecture as illustrated in Fig. 1. In the F-RAN, the user and control planes decoupled, where macro base stations (MBSs) are mainly used to provide seamless coverage and execute the functions of control plane, while RRHs are deployed to provide high speed data rate for the packet traffic transmission in the user plane. Furthermore, to explore the potential of edge devices like RRHs and user equipments (UEs), substantial amount of storage, communication, control, configuration, measurement and management were put at the edge of the network. Thus, part of radio signal processing and radio resource management can be achieved locally. Due to the cache in the RRHs and UEs, techniques like device to device (D2D), local distributed coordination, and global C-RAN are available in the F-RAN. With adaptive techniques and cache in the edge devices, the burden of the fronthaul and BBU pool can be alleviated.

The paper is organized as follows. In Section II, the F-RAN system model is presented and a problem about joint mode selection and resource allocation is formulated, while in Section III problem is solved and some heuristic results are derived. In Section IV, simulations results are demonstrated and conclusions are drawn in Section V.

Throughout this paper, we use $\{x\}^+$ to denote $\max\{x, 0\}$, and $\mathbb{E}[x]$ to denote the expectation of the random variable $x$. The complex Gaussian distribution is represented by $\mathcal{CN}(\cdot, \cdot)$. $|\cdot|$ is the size of a set and $\|\cdot\|_2$ is the $\ell_2$-th norm of a vector.

## II. System Model and Problem Formulation

Consider the downlink of a D2D underlay OFMD F-RAN having $N$ subchannels. Suppose that the F-RAN is assumed to operate in the slotted time mode with the unit time slot $t \in \{0, 1, 2, \cdots\}$. There are total of $I$ video files

QSHINE 2015, August 19-20, Taipei, Taiwan

DOI 10.4108/eai.19-8-2015.2260167

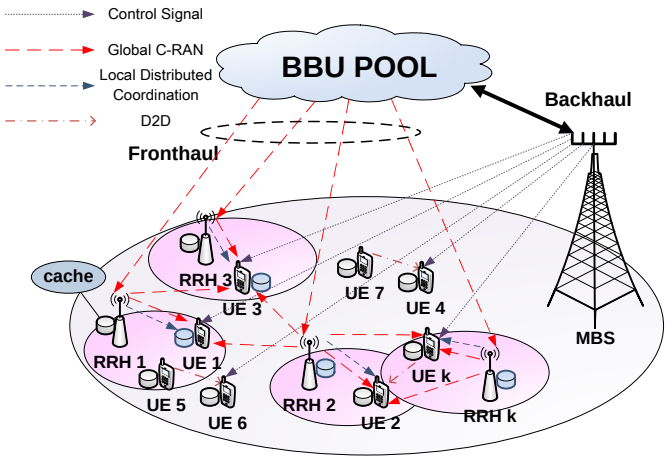


Fig. 1. The system model of a F-RAN. In the control plane the control signal is provided by the MBS while in the user plane, D2D, local distributed coordination, and global C-RAN can be chosen.

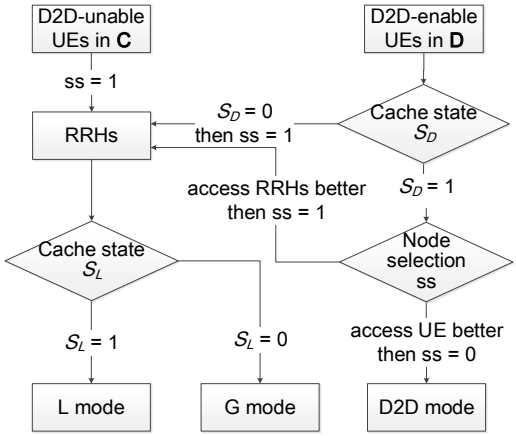


Fig. 2. An illustration of mode selection for UE $k$ in F-RAN at slot $t$ in subchannel $n$ under given cache state $S_L S_D$. Note that when $S_D = 1$, D2D pairs need to make a performance comparison between accessing RRHs and UEs.

in the cloud, where the size of the $i$-th video file is $F_i$ bits. There are a set of D2D pairs ($\mathbf{D}$) and a set of D2D-unable UEs ($\mathbf{C}$), which are uniformly labeled with $k = 1, 2, ..., K, K = |\mathbf{C}| + |\mathbf{D}| \leq N$ and access the files via $M$ RRHs equipped with $L$ antennas. The index of the video file requested by the $k$-th UE is denoted by $\pi_k \in \{1, ..., I\}$. Define $\boldsymbol{\pi} = \{\pi_1, ..., \pi_K\}$ as the user request profile (URP), which remains constant within a relatively long period. Let $\mathbf{h}_{k,n}(t) = [\mathbf{h}_{1,k,n}(t), ..., \mathbf{h}_{m,k,n}(t), ..., \mathbf{h}_{M,k,n}(t)] \in \mathbb{C}^{1\times ML}$ denote the channel state information (CSI) from all RRHs' transmit antennas to the UE $k$ in subchannel $n$ ($n = 1, 2, ..., N$), where $\mathbf{h}_{m,k,n}(t) \in \mathbb{C}^{1\times L}$ denotes the CSI matrix from $m$-th RRH's transmit antennas to the UE $k$ in subchannel $n$. $g_{j,k,n}(t) \in \mathbb{C}$ denotes the CSI from the UE $j$ to the UE $k$ in subchannel $n$ at slot $t$ where $j, k \in \mathbf{C} \cup \mathbf{D}$ and $m = 1, 2, ..., M$. Both $\mathbf{h}_{m,k,n}(t)$ and $g_{j,k,n}(t)$ are constant within a time slot.

### A. Maximum Distance Separable-Coded Random Caching

The video files are divided into segments and encoded using an ideal maximum distance separable rateless code. Each segment of $i$-th file is cached partially in the RRH, and the cache control variable $q_i^{\{1\}}$ is denoted as the proportion of cached bits in the segment [7]. As for UE, we assume the cache control variable as $q_i^{\{0\}}$, similarly. For simplicity, the contents in all RRH (UE) caches are assumed to be identical and the total cache data in RRH (UE) should be under the cache capacity of RRH $B_C^{\{1\}}$ (UE $B_C^{\{0\}}$).

$$\text{C1}: \quad \sum_{i=1}^{I} q_i^{\{0\}} F_i \leq B_C^{\{0\}},$$

$$\text{C2}: \quad \sum_{i=1}^{I} q_i^{\{1\}} F_i \leq B_C^{\{1\}}.$$

In F-RAN, there are two possible accessing nodes and three transmission modes. The impact of caching at RRHs and UEs on the physical layer are summarized by the cache state $S_L \in \{0, 1\}$ and $S_D \in \{0, 1\}$, respectively. The key to take advantage of the cache is increasing the probability of Local Distributed Coordination mode (L mode) and D2D mode in the system, which means the transmission strategy should align the transmissions of the cached data for different UEs as much as possible. Specifically, for given cache control vector $\mathbf{q} = [q_1^{\{0\}}, ..., q_I^{\{0\}}, q_1^{\{1\}}, ..., q_I^{\{1\}}]$ and URP $\boldsymbol{\pi}$, let $q_{\min}^{\{1\}} = \min_{1\leq k\leq K} \{q_{\pi_k}^{\{1\}}\}$ and $q_{\min}^{\{0\}} = \min_{1\leq k\leq K} \{q_{\pi_k}^{\{0\}}\}$. Conditioned on above, the cache state is generated using the following method. First, intermediate binary variables $\tilde{S_L}$ and $\tilde{S_D}$ are generated at each time slot according to a Bernoulli distribution with $Pr[\tilde{S_L} = 1] = q_{\min}^{\{1\}}$ and $Pr[\tilde{S_D} = 1] = q_{\min}^{\{0\}}$. If $\tilde{S_L} = 1$ ($\tilde{S_D} = 1$) and UEs have not received all the cached parity bits for the current requested segment, set $S_L(t) = 1$ ($S_D(t) = 1$). Otherwise, set $S_L(t) = 0$ ($S_D(t) = 0$).

### B. Cache-induced Mode Selection

We summarized the serving node selection by the selection state $ss \in \{0, 1\}$. Obviously, for D2D-unable UEs, its serving node is always RRH ($ss = 1$), while for D2D pairs, serving node may be RRH or UE in the same D2D pair ($ss \in \{0, 1\}$). Define the binary variable $x_{k,n}^{\{ss\}}(t)$ as the serving node selection and subchannel assignment indicator, with the value 1 if UE $k$ is assigned subchannel $n$ with selection state $ss$ at slot $t$, and 0 otherwise. Note that in each subchannel, at most one D2D-unable UE or one D2D pair can access the RRHs and at most one D2D pair can select D2D mode, which can be generalized into multiple UEs in one subchannel, and each UE can exploit subchannel $n$ by accessing at most one kind of nodes, which can be summarized as constrains C3, C4 and C5.

$$\text{C3}: \quad \sum\nolimits_{k=1}^{K} x_{k,n}^{\{0\}}(t) \leq 1 \qquad \forall n,$$

$$\text{C4}: \quad \sum\nolimits_{k=1}^{K} x_{k,n}^{\{1\}}(t) \leq 1 \qquad \forall n,$$

$$\text{C5}: \quad \sum\nolimits_{ss=0}^{1} x_{k,n}^{\{ss\}}(t) \leq 1 \qquad \forall k, n.$$

For UE $k$ accessing RRH in subchannel $n$ at slot $t$

($x_{k,n}^{\{1\}}(t) = 1$), there are two possible transmission modes, L mode when the request data is in the cache of the RRHs $S_L = 1$, and global C-RAN mode (G mode) when $S_L = 0$. When UE $k$ accesses UE in the same D2D pair in subchannel $n$ at slot $t$ ($x_{k,n}^{\{0\}}(t) = 1$), D2D mode is used and $S_D = 1$. Note that only when the profit of D2D mode is more and $S_D = 1$, $ss = 0$ and D2D could be used. Based on selection state $ss$ and cache state $S_L S_D$, the transmission mode selection is shown in Fig. 2 and the corresponding three possible transmission modes are:

- **D2D Mode**: In this mode, $ss = 0, S_D = 1$, the UE $k$ can be served as a receiver using D2D transmission, where the transmitter of D2D pair $k$ transmits a signal $v_{k,n}(t)s_{k,n}(t) \in \mathbb{C}$ to UE $k$. $v_{k,n}(t) \in \mathbb{C}$ is transmission precoding of D2D pair $k$ in subchannel $n$, and $s_{k,n}(t) \sim \mathcal{CN}(0,1)$ is the scalar-valued data stream for UE $k$ in subchannel $n$.
- **L Mode**: In this mode, $ss = 1, S_L = 1$, and the UE $k$ can only access its neighboring RRH $m$ in subchannel $n$. Consider linear precoding and minimum mean square error (MMSE) receiving for interference cancelation. For UE $k$, the serving RRH $m$ transmits a signal vector $\mathbf{v}_{m,k,n}(t)s_{k,n}(t) \in \mathbb{C}^{L\times 1}$, where $\mathbf{v}_{m,k,n}(t) \in \mathbb{C}^{L\times 1}$ denotes the precoding vector from RRH $m$ for UE $k$.
- **G Mode**: In this mode, $ss = 1, S_L = 0$, and the UE $k$ is served using coordinated multipoint between the RRHs in subchannel $n$. Consider linear precoding and MMSE receiving for interference cancelation similarly. In this case, the RRHs jointly transmit a signal vector $\mathbf{v}_{k,n}(t)s_{k,n}(t) \in \mathbb{C}^{ML\times 1}$ to UE $k$, where $\mathbf{v}_{k,n}(t) = [\mathbf{v}_{1,k,n}^T(t), \mathbf{v}_{2,k,n}^T(t), \ldots, \mathbf{v}_{M,k,n}^T(t)]^T \in \mathbb{C}^{ML\times 1}$ denotes the precoding vector for UE $k$ in subchannel $n$.

Note that the precoding vector $\mathbf{v}_{m,k,n}(t)$ and CSI $\mathbf{h}_{m,k,n}(t)$ can be respectively expressed as $\mathbf{v}_{m,k,n}(t) = \mathbf{D}_m\mathbf{v}_{k,n}(t)$ and $\mathbf{h}_{m,k,n}(t) = \mathbf{h}_{k,n}(t)\mathbf{D}_m^H$, where $\mathbf{D}_m = \{\underbrace{\mathbf{0}_L, \ldots, \mathbf{0}_L}_{m-1}, \mathbf{I}_L, \ldots, \mathbf{0}_L\} \in \mathbb{C}^{L\times ML}(m > 0)$. The L mode can be seen as a special case of G mode where there is only one RRH, namely RRH $m$ was allocated to power. Define the effective precoding vector for UE $k$ in subchannel $n$ as $\mathbf{S}_m(t)\mathbf{v}_{k,n}(t)$, where $\mathbf{S}_m(t) = S_L(t)\mathbf{D}_m^H\mathbf{D}_m + (1 - S_L(t))\mathbf{I}_{ML}$. The received signal at UE $k$ accessing RRHs in subchannel $n$ can be expressed as,

$$y_{k,n}^{\{1\}}(t) = \mathbf{h}_{k,n}(t)\mathbf{S}_m(t)\mathbf{v}_{k,n}(t)s_{k,n}(t) + z_{k,n}(t) + \sum_{j=1,j\neq k}^{K} x_{j,n}^{\{0\}}(t)g_{j,k,n}(t)v_{j,n}(t)s_{j,n}(t),$$

where $z_{k,n}(t) \sim \mathcal{CN}(0,\sigma^2)$ denotes the AWGN noise at the RUE $k$ in subchannel $n$. And the MMSE receiver for UE $k$ in subcarrier $n$ is,

$$u_{k,n}^{\{1\}}(t) = \mathbf{h}_{k,n}(t)\mathbf{S}_m(t)\mathbf{v}_{k,n}(t) \{\|\mathbf{h}_{k,n}(t)\mathbf{S}_m(t)\mathbf{v}_{k,n}(t)\|_2^2 + \Omega_{k,n}^{\{1\}}(t)\}^{-1},$$

where $\Omega_{k,n}^{\{1\}}(t)\} = \sum_{j=1,j\neq k}^{K} x_{j,n}^{\{0\}}(t)\|g_{j,k,n}(t)v_{k,n}(t)\|_2^2 + \sigma^2$.

The rate for UE $k$ accessing the RRHs is,

$$R_k^{\{1\}}(t) = \sum_{n=1}^{N} R_{k,n}^{\{1\}}(t) = \frac{B_0}{\ln 2}\sum_{n=1}^{N}\log(1 + \frac{x_{k,n}^{\{1\}}(t)\|\mathbf{h}_{k,n}(t)\mathbf{S}_m(t)\mathbf{v}_{k,n}(t)\|_2^2}{\Omega_{k,n}^{\{1\}}}(t)\}),$$

where $B_0$ is subchannel bandwidth.

Similarly, the received signal for UE $k$ accessing the UE in subchannel $n$ is,

$$y_{k,n}^{\{0\}}(t) = g_{k,k,n}(t)v_{k,n}(t)s_{k,n}(t) + z_{k,n}(t) + \sum_{j=1,j\neq k}^{K} x_{j,n}^{\{1\}}(t)\mathbf{h}_{k,n}(t)\mathbf{S}_m(t)\mathbf{v}_{j,n}(t)s_{j,n}(t).$$

And the MMSE receiver for UE $k$ in subcarrier $n$ is,

$$u_{k,n}^{\{0\}}(t) = g_{k,k,n}(t)v_{k,n}(t) \{\|g_{k,k,n}(t)v_{k,n}(t)\|_2^2 + \Omega_{k,n}^{\{0\}}(t)\}^{-1},$$

where $\Omega_{k,n}^{\{0\}}(t) = \sum_{j=1,j\neq k}^{K} x_{j,n}^{\{1\}}(t)\|\mathbf{h}_{k,n}(t)\mathbf{S}_m(t)\mathbf{v}_{j,n}(t)\|_2^2 + \sigma^2$.

The corresponding rate for D2D pair $k$ is,

$$R_k^{\{0\}}(t) = \sum_{n=1}^{N} R_{k,n}^{\{0\}}(t) = \frac{B_0}{\ln 2}\sum_{n=1}^{N}\log(1 + \frac{x_{k,n}^{\{0\}}(t)\|g_{k,k,n}(t)v_{k,n}(t)\|_2^2}{\Omega_{k,n}^{\{0\}}(t)\}}).$$

The power consumption for RRH $m$ and UE $k$ are,

$$P_m^{\{1\}}(t) = \sum_{n=1}^{N} P_{m,n}^{\{1\}}(t) = \sum_{n=1}^{N}\sum_{k=1}^{K} x_{k,n}^{\{1\}}\|\mathbf{D}_m\mathbf{v}_{k,n}(t)\|_2^2,$$

$$P_k^{\{0\}}(t) = \sum_{n=1}^{N} P_{k,n}^{\{0\}}(t) = \sum_{n=1}^{N} x_{k,n}^{\{0\}}\|v_{k,n}(t)\|_2^2,$$

respectively. The energy efficiency (EE) is defined as,

$$\eta(t) = \sum_{n=1}^{N}\{\frac{\alpha}{K}\sum_{k=1}^{K} R_{k,n}^{\{0\}}(t) + \frac{\alpha}{K}\sum_{k=1}^{K} R_{k,n}^{\{1\}}(t) - \frac{\beta}{K}\sum_{k=1}^{K} P_{k,n}^{\{0\}}(t) - \frac{\beta}{M}\sum_{m=1}^{M} P_{m,n}^{\{1\}}(t)\},$$

where $\alpha, \beta$ are weighting factor. Note that the solution to the optimization problem about EE $\eta(t)$, is also a Pareto optimal solution to problem about EE in traditional form and can be used to characterize the EE in traditional form [8].

### C. Problem Formulation

Suppose there are queues maintained for UEs in F-RANs as illustrated in Fig. 3 which are represented by $\mathbf{H}(t) = \{H_k(t)|\, k = 1, ..., K\}$, where $H_k(t)$ denotes the queue backlog for the UE $k$ at the time slot $t$. The random traffic arrival for the RUE $k$ at the time slot $t$ is denoted by $A_k(t)$, which is assumed to be independent and identically distributed over time slots. Define the set of $A_k(t)$ as $\mathbf{A}(t) = \{A_k(t)|k = 1, ..., K\}$, and the arrival rates of queues are $\lambda = \mathbb{E}\{A_k(t)\}$.

At each time slot, the arrival and departure rates of the UE $k$ are $A_k(t)$ and $R_k(t) = R_k^{\{1\}}(t) + R_k^{\{0\}}(t)$, respectively. Therefore, $H_k(t)$ evolves according to $H_k(t+1) = (H_k(t) - R_k(t))^+ + A_k(t)$.

Considering the random and bursty characteristics of traffic arrivals and the quality of service requirement of RUEs in F-RANs, it is imperative to consider delay-aware resource allocation techniques. Therefore, to achieve this objective, the queue stability is first defined.

*Definition 1: A discrete time process $Q(t)$ is mean rate stable [9] if*

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

Denote the whole precoding as $\mathcal{V}(t) = \{\mathbf{v}_{k,n}(t), v_{k,n}(t) : \forall n, k\}$, and the whole serving node selection and subchannel assignment indicator as $\mathcal{X}(t) = \{x_{k,n}^{\{ss\}}(t) : \forall n, k, ss\}$, the joint mode selection and resource allocation problem (M-SRAP) can be described as follows:

$$\begin{aligned}
&\max_{\{\mathbf{q},\mathcal{X}(t),\mathcal{V}(t)\}} \quad \mathbb{E}\{\eta(t)\} \\
&s.t. \quad \text{C1,C2,C3,C4,C5,} \\
&\text{C6}: \quad x_{k,n}^{\{0\}}(t) = 0 \qquad \forall k \in \mathbf{C}, n, \\
&\text{C7}: \quad x_{k,n}^{\{ss\}}(t) \in \{0,1\} \qquad \forall k, n, ss, \\
&\text{C8}: \quad \mathbf{H}(t) \text{ are mean rate stable,} \\
&\text{C9}: \quad \mathbb{E}\{P_k^{\{0\}}(t)\} \le \tilde{P}_{\text{thres}}^{\{0\}} \qquad \forall k \in \mathbf{D}, \\
&\text{C10}: \quad \mathbb{E}\{P_m^{\{1\}}(t)\} \le \tilde{P}_{\text{thres}}^{\{1\}} \qquad \forall m \in \{1, \ldots, M\}.
\end{aligned} \tag{1}$$

Problem (1) is about the average EE maximization of the F-RAN under constraints. C6 and C7 are definition constraints which have been mentioned above. C8 is used to guarantee a finite queue's length. C9 and C10 are average and instantaneous power consumption constraints for RRH $m$ and UE $k$.

## III. Joint Mode Selection and Resource Allocation

The MSRAP (1) can be decomposed into two subproblems:

- Short-term mode selection and resource allocation $\mathcal{X}(t), \mathcal{V}(t)$:

$$\begin{aligned}
&\max_{\{\mathcal{X}(t),\mathcal{V}(t)\}} \quad \mathbb{E}\{\eta(t)\} \\
&s.t. \quad \text{C3} \sim \text{C10}.
\end{aligned} \tag{2}$$

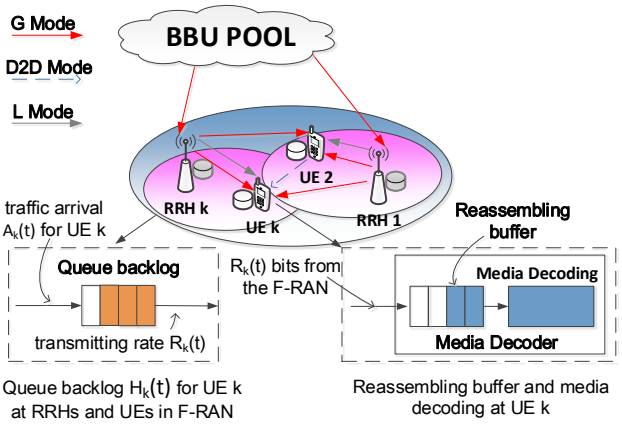


Fig. 3. An illustration of queues' dynamics in F-RAN, including two parts as illustrated in the picture. Once the total number of parity bits for current segment at the reassembling buffer is equal to the length of one segment, the whole segment is decoded at the media decoder and the reassembling buffer is cleared so that the media decoder can read the next segment from the playback buffer.

- Long-term cache control $\mathbf{q}$ for given $\mathcal{X}(t), \mathcal{V}(t)$:

$$\begin{aligned}
&\max_{\{\mathbf{q}\}} \quad \mathbb{E}\{\eta(t)\} \\
&s.t. \quad \text{C1,C2.}
\end{aligned} \tag{3}$$

Due to the space limitations, the solution to only subproblem (2) is given. Construct virtue queues $\mathbf{F}(t) = \{F_k(t)|\, k = 1, ..., K\}$ and $\mathbf{G}(t) = \{G_m(t)|\, m = 1, ..., M\}$, where $F_k(t+1) = (F_k(t) - P_{\text{thres}}^{\{0\}})^+ + P_k^{\{0\}}(t)$ and $G_m(t+1) = (G_m(t) - P_{\text{thres}}^{\{1\}})^+ + P_m^{\{1\}}(t)$. Suppose $\mathbb{E}\{F_k(0)\} < \infty$ and $\mathbb{E}\{G_m(0)\} < \infty$. If virtual queue $F_k(t)$ and $G_m(t)$ are mean rate stable, C9 and C10 can be satisfied [9].

Denote $\mathbf{\Theta}(t) = [\mathbf{H}(t), \mathbf{F}(t), \mathbf{G}(t)]$ as the combined matrix of all the queues. The Lyapunov function is defined as a scalar metric of queue congestion [9],

$$L(\mathbf{\Theta}(t)) \triangleq \frac{1}{2}\left\{\sum_{k=1}^{K} H_k^2(t) + \sum_{k=1}^{K} F_k^2(t) + \sum_{m=1}^{M} G_m^2(t)\right\}.$$

And the Lyapunov drift [9] is defined as,

$$\Delta(\mathbf{\Theta}(t)) \triangleq \mathbb{E}\{L(\mathbf{\Theta}(t+1)) - L(\mathbf{\Theta}(t))\}.$$

Thus the subproblem (2) can be solved by choosing proper mode and resource allocation at every time slot to minimize the supremum of the drift plus penalty $\Delta(\mathbf{\Theta}(t)) - V\mathbb{E}\{\eta(t)|\mathbf{\Theta}(t)\}$ given $\mathbf{\Theta}(t)$. Based on the *Lemma 4.6* in [9] and the concept of opportunistically minimizing an expectation, the subproblem (2) has become a new MSRAP as follows:

$$\begin{aligned}
&\max_{\{\mathcal{X}(t),\mathcal{V}(t)\}} \sum_{n=1}^{N}\Big\{\sum_{k=1}^{K}\alpha_k R_{k,n}^{\{0\}}(t) + \sum_{k=1}^{K}\alpha_k R_{k,n}^{\{1\}}(t) \\
&\qquad - \sum_{k=1}^{K}\beta_k P_{k,n}^{\{0\}}(t) - \sum_{m=1}^{M}\gamma_m P_{m,n}^{\{1\}}(t)\Big\} \\
&s.t. \quad \text{C3,C4,C5,C6,C7,}
\end{aligned} \tag{4}$$

where $\alpha_k = H_k(t) + \frac{V\alpha}{K}, \beta_k = F_k(t) + \frac{V\beta}{K}, \gamma_m = G_m(t) + \frac{V\beta}{M}$.

The MSRAP (4) can be solved via block coordinate descent method by iterating among $\mathcal{X}(t), \mathcal{V}(t)$:

- The optimization problem for finding the optimal $\mathcal{X}(t)$ under the fixed $\mathcal{V}(t)$ is,

$$\begin{aligned}\max_{\{\mathcal{X}(t)\}} \sum_{n=1}^{N}\Big\{ \sum_{k=1}^{K} \alpha_k R_{k,n}^{\{0\}}(t) + \sum_{k=1}^{K} \alpha_k R_{k,n}^{\{1\}}(t) \\ - \sum_{k=1}^{K} \beta_k P_{k,n}^{\{0\}}(t) - \sum_{m=1}^{M} \gamma_m P_{m,n}^{\{1\}}(t)\Big\} \\ s.t. \quad \text{C3,C4,C5,C6,C7.}\end{aligned} \tag{5}$$

  The problem (5) can be solved based on particle swarm optimization [10] as follows: Assume there are $B$ particles, and the velocity and position of particles are respectively $\mathbf{v}_b = (v_b^1, v_b^2, \ldots, v_b^{2N})$ and $\mathbf{x}_b = (x_b^1, x_b^2, \ldots, x_b^{2N})$, which are updated according to,

$$\begin{aligned}\mathbf{v}_b^{u+1} &= w\mathbf{v}_b^u + r_1 c_1(\mathbf{p}_b^u - \mathbf{x}_b^u) + r_2 c_2(\mathbf{g}^u - \mathbf{x}_b^u), \\ \mathbf{x}_b^{u+1} &= \mathbf{v}_b^{u+1} + \mathbf{x}_b^u,\end{aligned} \tag{6}$$

  where $u = 1, 2, ..., U$ denotes the $u$-th iteration, $\mathbf{p}_b$ is the personal best position which has been found by the particles so far, and $\mathbf{g}$ is the global best position which has been found by the whole particle swarm so far.
  The mapping between $x_b^n$ and $x_{k,n}^{\{ss\}}(t)$ is,

$$x_{k,n}^{\{ss\}}(t) = \begin{cases} 1 & \text{if } k = \lfloor (K+1)x_b^n \rfloor, ss = 1, x_b^n \in (0,1), \\ 1 & \text{if } k = S_D(t)\left\lceil (|\mathbf{D}|+1)x_b^{N+n} + |\mathbf{C}| \right\rceil, \\ & \quad ss = 0, x_b^{N+n} \in (0,1), \\ 0 & \text{others.} \end{cases} \tag{7}$$

  And note that when $\lfloor (K+1)x_b^n \rfloor = 0$, it means subchannel $n$ isn't assigned to accessing to RRH, when $\left\lceil (|\mathbf{D}|+1)x_b^{N+n} \right\rceil + |\mathbf{C}| = K+1$, it means subchannel $n$ isn't assigned to D2D mode. Specially, $k = \lfloor (K+1)x_b^n \rfloor = \left\lceil (|\mathbf{D}|+1)x_b^{N+n} \right\rceil + |\mathbf{C}|$ means that UE $k$ accesses both UE and RRHs in subchannel $n$, which violates C5. Thus $x_{k,n}^{\{0\}}(t)$ is forced to be 1, while $x_{k,n}^{\{1\}}(t)$ is forced to be 0.
- The optimization problem for finding the optimal $\mathcal{V}(t)$ under the fixed $\mathcal{X}(t)$. In a dedicated subchannel $n_0$, there are at most two different UEs due to constraints. Suppose there are two UEs in one subchannel, for example $x_{k_0,n_0}^{\{0\}} = x_{k_1,n_0}^{\{1\}} = 1, k_0 \neq k_1$. It becomes

$$\begin{aligned}\max_{\{\mathcal{V}(t)\}} \alpha_{k_0} R_{k_0,n_0}^{\{0\}}(t) + \alpha_{k_1} R_{k_1,n_0}^{\{1\}}(t) \\ -\beta_{k_0}\|v_{k_0,n_0}(t)\|_2^2 - \sum_{m=1}^{M} \gamma_m \|\mathbf{D}_m \mathbf{v}_{k_1,n_0}(t)\|_2^2.\end{aligned} \tag{8}$$

  The problem (8) has the same optimal solution as the following WMMSE minimization problem [11]:

$$\begin{aligned}\min_{\{\omega_{k,n_0}^{\{ss\}}, u_{k,n_0}^{\{ss\}}, \mathcal{V}(t)\}} \alpha_{k_0}\{\omega_{k_0,n_0}^{\{0\}} e_{k_0,n_0}^{\{0\}} - \log \omega_{k_0,n_0}^{\{0\}}\} \\ + \alpha_{k_1}\{\omega_{k_1,n_0}^{\{1\}} e_{k_1,n_0}^{\{1\}} - \log \omega_{k_1,n_0}^{\{1\}}\} \\ + \beta_{k_0}\|v_{k_0,n_0}(t)\|_2^2 \\ + \sum_{m=1}^{M} \gamma_m \|\mathbf{D}_m \mathbf{v}_{k_1,n_0}(t)\|_2^2,\end{aligned} \tag{9}$$

  where $\omega_{k,n_0}^{\{ss\}}$ denotes the mean-square error (MSE) weight for UE $k \in \{k_1, k_2\}$ in sunchannel $n_0$, $u_{k,n_0}^{\{ss\}}$ is the MMSE receiver and $e_{k,n_0}^{\{ss\}}$ is the the corresponding MSE defined as,

$$e_{k,n}^{\{ss\}}(t) \triangleq \mathbb{E}\{(u_{k,n}^{\{ss\}}(t) y_{k,n}^{\{ss\}}(t) - s_{k,n}(t))^2\}.$$

  Since problem (9) is convex in each of the optimization variables $\omega_{k,n_0}^{\{ss\}}, u_{k,n_0}^{\{ss\}}, \mathcal{V}(t)$, it can be solved via the block coordinate descent method. First, for fixed $u_{k,n_0}^{\{ss\}}, \mathcal{V}(t)$, the optimal $\omega_{k,n_0}^{\{ss\}}$ is given by $\omega_{k,n_0}^{\{ss\}} = \frac{1}{e_{k,n}^{\{ss\}}(t)}$. Second, for fixed $\omega_{k,n_0}^{\{ss\}}, \mathcal{V}(t)$, the optimal $u_{k,n_0}^{\{ss\}}$ is given by the MMSE receiver above. Finally, for fixed $\omega_{k,n_0}^{\{ss\}}, u_{k,n_0}^{\{ss\}}$, problem (9) is a quadratically constrained quadratic optimization problem which can be solved.

Our proposed scheme is described as follows:

**Algorithm 1** The MSRAP scheme at slot $t$.

**Require:** Iteration number $U$, particles number $B$, velocity and position of particles $\mathbf{v}_b^0$ and $\mathbf{x}_b^0$, personal best position $\mathbf{p}_b^0$, global best position $\mathbf{g}^0$, precoding $\mathcal{V}(t)$cache state $S_L(t)S_D(t)$;
**Ensure:** Mode selection and resource allocation $x_{k,n}^{\{ss\}}(t)$, $\mathcal{V}(t)$.
  **for** $u = 1 : U$ **do**
    **for** $b = 1 : B$ **do**
      Convert $\mathbf{x}_b$ to $\mathcal{X}(t)$;
      Decide precoding $\mathcal{V}(t)$;
      Calculate the fitness value according to (4);
    **end for**
    Update the personal best position of each particle $\mathbf{p}_b^u$ and the global best position $\mathbf{g}^u$ according to the fitness value;
    Update the velocity $\mathbf{v}_b^{u+1}$ and position of each particle $\mathbf{x}_b^{u+1}$;
  **end for**

## IV. Simulation Results and Analysis

Consider a F-RAN with parameters shown in Table I. The number of particles $B = 5$, the maximum number of iterations $U = 50$, the learning factors $c_1 = c_2 = 2$, the inertia weight $w$ decreasing linearly from 0.9 to 0.4.

Fig. 4 shows that the average EE of systems versus the control parameter $V$ for different cache control variables,

TABLE I
SIMULATION PARAMETERS

| | |
|---|---|
| Num. of RRHs $M$, num. of RRH antennas $L$ and subchannels $N$ | (5, 2, 5) |
| Size of set $\mathbf{D}$, set $\mathbf{C}$ | 2, 3 |
| Path loss exponent for transmission | 4 |
| Noise power spectral density | -174 dBm/Hz |
| Subchannel bandwidth | 0.2 MHz |
| Small-scale fading | Rayleigh fading |
| Average transmit power constraint of RRH, UE | 500 mW, 200mW |

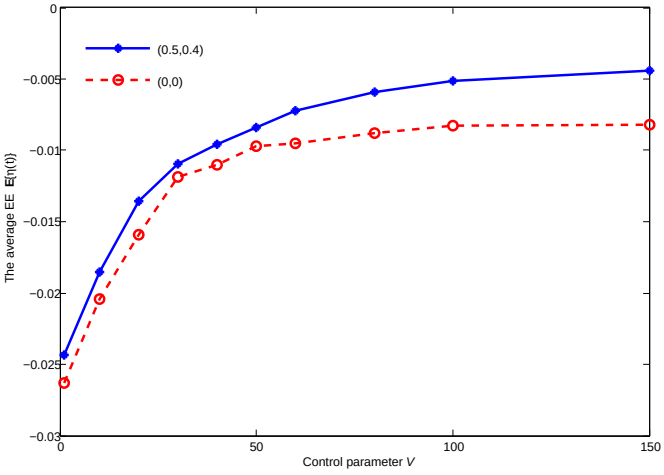


Fig. 4. The average EE of systems versus the control parameter $V$ under different cache control variables $(q_{\min}^{\{1\}}, q_{\min}^{\{0\}})$.

where the average EE performance increases with $V$ till saturation. When $(q_{\min}^{\{1\}}, q_{\min}^{\{0\}}) = (0, 0)$, G Mode is the only transmission mode can be chosen and the F-RAN has degenerated into a traditional C-RAN. It is shown that the F-RAN with cache owns a significant performance gain over traditional C-RAN as the incorporation of cache, especially when fronthaul consumptions were taken into account.

Fig. 5 shows that for given arrival rate, the average queue backlog grows linearly in $\mathcal{O}(V)$, which presents a tradeoff between average EE and average delay combined with Fig. 4. This is because a larger $V$ leads more emphasizes on EE at the cost of incurring worse queueing delays. Besides, it is also shown that the F-RAN outperforms the C-RAN in the aspect of delay due to the various transmission modes.

## V. CONCLUSION

In this paper, we proposed a new framework fog computing based radio access network (F-RAN) to explore the cache incorporation into cloud-RAN. The joint mode selection and resource allocation problem (MSRAP) in F-RANs supported device to device is studied, which maximizes the system energy efficiency (EE) under constraints about delay and resource reuse. The proper mode selection and resource allocation is also derived through solving the MSRAP and it is shown that the cache incorporation can significantly improve the system

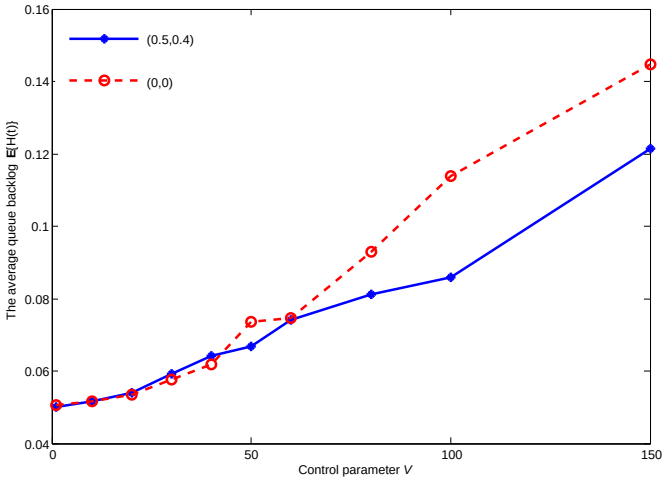


Fig. 5. The average queue backlog versus the control parameter $V$ under different cache control variables $(q_{\min}^{\{1\}}, q_{\min}^{\{0\}})$.

performance. Numerical results have validated the benefits of cache and an EE-delay tradeoff is finally achieved by the proposed algorithm.

## REFERENCES


[1] CMCC, "C-RAN the road towards green ran," *CMCC white paper*, Oct. 2011.

[2] A. Checko, H. L. Christiansen, Y. Ying, *et. al.*, "Cloud RAN for mobile networks-a technology overview," *IEEE Commun. Suevey & Tutorials*, vol. 17, no. 1, pp. 405-426, Sep. 2014.

[3] J. G. Andrews, S. Buzzi, C. Wan, *et. al.*, "What will 5G be," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065-1082, Jun. 2014.

[4] J. Erman, A. Gerber, M. Hajiaghyi, *et. al.*, "To cache or not to cache: The 3G case," *IEEE Internet Comput.*, vol. 15, no. 2, pp. 27-34, Mar. 2011.

[5] B. A. Ramanan, L. M. Drabeck, M. Haner, *et. al.*, "Cacheability analysis of HTTP traffic in an operational LTE network," *Wireless Telecommun. Symp.*, Apr. 2013.

[6] X. Wang, M. Chen, T. Taleb, *et. al.*, "Cache in the air: Exploiting content caching and delivery techniques for 5G systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 131-139, Feb. 2014.

[7] A. Liu and V. Lau, "Exploiting base station caching in MIMO cellular networks: Opportunistic cooperation for video streaming," *IEEE Trans. Signal Process.*, vol. 63, no. 1, pp. 57-69, Jan. 2015.

[8] C. He, B. Sheng, P. Zhu, *et. al.*, "Energy- and spectral-efficiency tradeoff for distributed antenna systems with proportional fairness," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 5, pp. 894-902, May 2013.

[9] M. Neely, *Stochastic Network Optimization with Application to Communication and Queuing Systems*. Morgan & Claypool, 2010.

[10] Y. Gong, J. Zhang, H. S. Chung, *et. al.*, "An efficient resource allocation scheme using particle swarm optimization," *IEEE Trans. Evol. Comput.*, vol. 16, no. 6, pp. 801-816, Dec. 2012.

[11] Q. Shi, M. Razaviyayn, Z. Luo, and H. Chen, "An iteratively weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4331-4340, Sep. 2011.