Adapting Downlink Power in Fronthaul-Constrained Hierarchical Software-Defined RANs

Xianfu Chen, Zhu Han, Zheng Chang, Guoliang Xue, Honggang Zhang, and Mehdi Bennis

***Abstract*—The proof-of-concept software-defined radio access network (RAN) is not flexible enough due to the inherent delay and the necessity of high-capacity fronthaul links. We are hence motivated to propose a hierarchical software-defined RAN architecture, over which the base stations (BSs) are abstracted into multiple virtual local controllers while these local controllers are administered by a high-level controller. Under such a hierarchical network architecture, we particularly investigate in this paper how to adapt the BS transmit power over a long term according to the network dynamics under the constraints of mobile user queue stability and limited fronthaul capacity. We first formulate an off-line stochastic power adaptation problem. Through developing the Lyapunov method, we transform the problem into an approximate on-line optimization task. However, the challenge arises from the introduced per-cluster fronthaul capacity constraint. To solve the task efficiently and avoid extensive information exchange between the high-level controller and the local controllers, we put forward a novel low-complexity algorithm by designing a non-cooperative power adaptation game among the local controllers. Simulations are provided to evaluate the efficacy of the proposed studies.**

I. Introduction

The exponentially growing mobile data traffic leads to the need of ever increasing capacity density in radio access networks (RANs) [1]. To keep pace with such demands, one of the promising solutions is to make the network infrastructure heterogeneous and dense. In a dense environment, the base stations (BSs) have to be operated over a common spectrum band, making the network operations extremely complex due to the tight coupling of control plane decisions at the neighbouring BSs. Moreover, the traditional RANs are dimensioned to cope with the peak traffic demands. Such designs are not flexible enough to match the radio resources with the spatially and temporally fluctuating traffics, resulting in low spectral efficiency and inferior energy efficiency as well [2].

Applying the idea of software-defined networking to RANs brings the immediate advantages of simplifying the management of a dense network. In a software-defined RAN, the control plane is decoupled from the data plane via virtualizing all independent BSs as a controller which makes centralized control plane decisions [3]. The centralized controller can thus optimize the network performance with the global view of the network and adapt radio resources to the network dynamics, i.e., the channel quality variations and the mobile traffic fluctuations from mobile users (MUs). However, the problems with the software-defined RAN concept lie in the inherent latency due to the totally centralized control plane and the need of large-capacity fronthaul links to connect the BSs and the controller [4], [5]. It thus becomes necessary to design a more flexible software-defined RAN architecture.

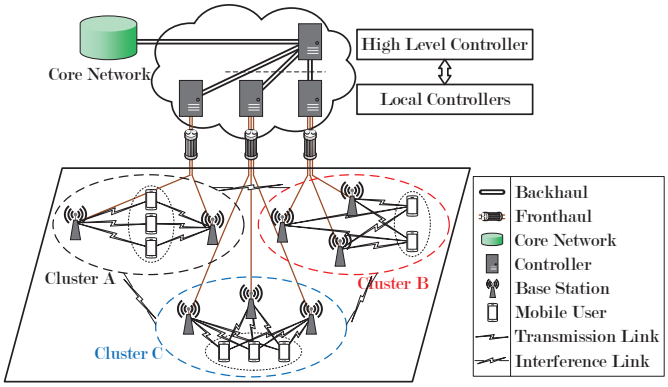

Fig. 1. Illustrative example of a hierarchical software-defined radio access network.

In this paper, we first consider a new hierarchical architecture of the software-defined RAN, which is shown in Fig. 1. Instead of virtualizing all BSs as a single centralized controller, multiple clusters are formed with regards to the BS geographic locations, with each being assigned a virtual local controller. The connections between the clusters and their associated local controllers are established via the capacity-limited fronthaul links [6]. A virtual high-level controller is responsible for coordinating control plane decisions among the local controllers. With the hierarchical design, the global network view at the high-level controller is aggregated across the local clusters. The local controllers make local decisions (e.g., intra-cluster interference mitigation), which avoids distant control from the high-level controller and thus alleviates control latency.

In spite of the benefits, challenges remain in facilitating a hierarchical software-defined RAN (HSDRAN) in practice. Particularly, mechanisms that efficiently utilize the decoupling of control plane and data plane in a hierarchical network architecture and the capacity-limited fronthaul links must be

X. Chen is with the VTT Technical Research Centre of Finland, Finland (e-mail: xianfu.chen@vtt.fi). Z. Han is with the Department of Electrical and Computer Engineering as well as the Department of Computer Science, University of Houston, Houston, TX, USA (e-mail: zhan2@uh.edu). Z. Chang is with the Department of Mathematical Information Technology, University of Jyväskylä, Finland (e-mail: zheng.chang@jyu.fi). G. Xue is with the Ira A. Fulton Schools of Engineering, Arizona State University, Tempe, AZ, USA (e-mail: xue@asu.edu). H. Zhang is with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China (e-mail: honggangzhang@zju.edu.cn). M. Bennis is with the Centre for Wireless Communications, University of Oulu, Finland (e-mail: bennis@ee.oulu.fi).

developed to achieve stable data transmissions for MUs. There exists related work addressing part of the issues. In [7], the authors proposed a distributive stochastic learning algorithm for dynamic clustering to minimize the long-term average delay for MUs. The authors of [8], [9] developed resource allocation techniques to optimize the network performance considering the limited-capacity fronthaul constraints. The Lyapunov optimization framework has been adopted to achieve a suboptimal solution to a stochastic optimization constrained by the queue stability [10], [11]. We emphasize that the work in this paper is different from the literature from the following three perspectives. First, we optimize the performance from both the network and the MU sides by adapting the BS transmit powers. Second, we formulate an off-line stochastic optimization problem to minimize the long-term average transmit power accumulated over all BSs, under the constraints of limited fronthaul capacity and MU queue stability. Third, by developing the Lyapunov method, we transform the off-line problem to an on-line optimization problem, for which a low-complexity algorithm is proposed.

The rest of the paper is organized as follows. In Section II, we elaborate on details of the system model that is considered in this paper. In Section III, we introduce the long-term average transmit power minimization in a HSDRAN as an off-line stochastic optimization problem and present the approximate on-line version. For the formulated on-line optimizatiom, we propose a low-complexity algorithm in Section IV. In Section V, we provide preliminary simulation results to validate the proposed work. We conclude this paper in Section VI.

II. System Descriptions

As depicted in Fig. 1, we consider a HSDRAN, the heart of which constitutes a hierarchical structure of virtual controllers. The local controllers process baseband signals for MUs and optimize cluster-wide radio resource utilization. The analysis throughout this paper focuses on downlink communications. In the downlink, BSs communicate with MUs using RF signals, which carry the baseband signals and the precoding vectors that are determined by the local controllers. All BSs in the network share a common spectrum of bandwidth W (Hz). The time dimension is partitioned into discrete time slots indexed by $t \in \mathbb{N}^+$ with equal duration δ (seconds). Let $\mathcal{K}$ be the set of local controllers/clusters. In each cluster $k \in \mathcal{K}$, a set $\mathcal{B}_k$ of BSs serve a set $\mathcal{I}_k$ of single-antenna MUs. For notational convenience, we assume that every BS is equipped with N antennas.

We denote $\rho_{i_k}^{b_k}(t) \in \mathbb{R}_+$ and $\mathbf{u}_{i_k}^{b_k}(t) \in \mathbb{C}^{N\times 1}$, respectively, as the allocated transmit power and the precoding vector of BS $b_k \in \mathcal{B}_k$ for MU $i_k \in \mathcal{I}_k$ at each time slot t, where the precoding vector $\mathbf{u}_{i_k}^{b_k}(t)$ has unit norm, i.e., $\|\mathbf{u}_{i_k}^{b_k}(t)\| = 1$. The baseband signal $y_{i_k}(t) \in \mathbb{C}$ received by MU i_k at time t can thus be expressed by (1), where $\mathbf{h}_{i_k}^{b_{k'}}(t) \in \mathbb{C}^{N\times 1}$ denotes the vector of channels between antennas of BS $b_{k'}$ and MU i_k, $x_{i_k}(t) \in \mathcal{CN}(0,1)$ denotes the transmitted signal for MU i_k, $z_{i_k}(t) \in \mathcal{CN}(0,\sigma^2)$ denotes the additive white noise, and the superscript H denotes the Hermitian transpose operation. The channels hold their states for the duration of one time slot, and potentially change states at the time slot boundaries in an independent and identically distributed (i.i.d.) manner. Based on the instantaneous local cluster channel state information, i.e., $\mathbf{h}_k(t) = \{\mathbf{h}_{i_k}^{b_k}(t) : b_k \in \mathcal{B}_k, i_k \in \mathcal{I}_k\}$, local controller k adopts zero-forcing (ZF) precoding to cancel the intra-cluster interference. Specifically, the collection of local cluster ZF precoding vectors $\mathbf{u}_k(t) = \{\mathbf{u}_{i_k}^{b_k}(t) : b_k \in \mathcal{B}_k, i_k \in \mathcal{I}_k\}$ should satisfy: $(\mathbf{h}_{i_k}^{b_k}(t))^{\mathsf{H}}\mathbf{u}_{i'_k}^{b_k}(t) = 0$, if $i'_k \neq i_k$. Assuming that the transmitted signals and the noise are mutually independent, we derive the signal-to-interference-plus-noise ratio (SINR) at MU i_k during the time slot as follows

$$\mathrm{SINR}_{i_k}(t) = \frac{\left|\sum_{b_k\in\mathcal{B}_k}\sqrt{\rho_{i_k}^{b_k}(t)}\left(\mathbf{h}_{i_k}^{b_k}(t)\right)^{\mathsf{H}}\mathbf{u}_{i_k}^{b_k}(t)\right|^2}{M_{i_k}(t)+\sigma^2}, \quad (2)$$

where

$$M_{i_k}(t) = \sum_{k'\in\mathcal{K}\setminus\{k\}}\sum_{i_{k'}\in\mathcal{I}_{k'}}\left|\sum_{b_{k'}\in\mathcal{B}_{k'}}\sqrt{\rho_{i_{k'}}^{b_{k'}}(t)}\left(\mathbf{h}_{i_k}^{b_{k'}}(t)\right)^{\mathsf{H}}\mathbf{u}_{i_{k'}}^{b_{k'}}(t)\right|^2, \quad (3)$$

is the power of received inter-cluster interference. The corresponding data rate is given by

$$R_{i_k}(t) = W\log_2(1+\mathrm{SINR}_{i_k}(t)). \quad (4)$$

The transmit power consumed by a BS b_k during time slot t can thus be computed as

$$P_{b_k}(t) = \sum_{i_k\in\mathcal{I}_k}\rho_{i_k}^{b_k}(t). \quad (5)$$

At each local controller $k \in \mathcal{K}$, there are $|\mathcal{I}_k|$ separate downlink queues which buffer data bits for MUs in the cluster, where $|\mathcal{I}|$ is the cardinality of the set $\mathcal{I}$. Let $Q_{i_k}(t) \in \{0\}\cup\mathbb{N}^+$ be the queue length (number of bits) of a MU i_k at the beginning of each time t, its queue evolution can be expressed as in the form below,

$$Q_{i_k}(t+1) = [Q_{i_k}(t) - \delta R_{i_k}(t), 0]^+ + A_{i_k}(t), \quad (6)$$

where $[q]^+$ is the projection of q into a non-negative area and $A_{i_k}(t)$ is the number of data bits arriving at the end of time slot t. The data arrivals get queued until they are transmitted. Like $\mathbf{h}_{i_k}^{b_{k'}}(t)$, we assume that $A_{i_k}(t)$ is i.i.d. across time slots according to a general distribution $\mathsf{Pr}\{A_{i_k}(t)\}$.

Next, we shall address the data transmissions from local controllers to BSs over the fronthaul links. It's worth noting that in a cluster $k \in \mathcal{K}$, $\rho_{i_k}^{b_k}(t) = 0$ implies that BS $b_k \in \mathcal{B}_k$ does not serve MU $i_k \in \mathcal{I}_k$ during time slot t. In contrast, if $\rho_{i_k}^{b_k}(t) \neq 0$, the fronthaul link between local controller k and BS b_k needs to carry the baseband signal $x_{i_k}(t)$ for MU i_k.

$$y_{i_k}(t)=\underbrace{\left(\sum_{b_k\in\mathcal{B}_k}\sqrt{\rho_{i_k}^{b_k}(t)}\left(\mathbf{h}_{i_k}^{b_k}(t)\right)^{\mathsf{H}}\mathbf{u}_{i_k}^{b_k}(t)\right)x_{i_k}(t)}_{\text{the desired signal}}+\underbrace{\sum_{i'_k\in\mathcal{I}_k\setminus\{i_k\}}\left(\sum_{b_k\in\mathcal{B}_k}\sqrt{\rho_{i'_k}^{b_k}(t)}\left(\mathbf{h}_{i_k}^{b_k}(t)\right)^{\mathsf{H}}\mathbf{u}_{i'_k}^{b_k}(t)\right)x_{i'_k}(t)}_{\text{the intra-cluster interference}}$$
$$+\underbrace{\sum_{k'\in\mathcal{K}\setminus\{k\}}\sum_{i_{k'}\in\mathcal{I}_{k'}}\left(\sum_{b_{k'}\in\mathcal{B}_{k'}}\sqrt{\rho_{i_{k'}}^{b_{k'}}(t)}\left(\mathbf{h}_{i_k}^{b_{k'}}(t)\right)^{\mathsf{H}}\mathbf{u}_{i_{k'}}^{b_{k'}}(t)\right)x_{i_{k'}}(t)+z_{i_k}(t)}_{\text{the inter-cluster interference plus the noise}} \tag{1}$$

As a result, the allocated fronthaul capacity for BS b_k can be mathematically written as[1]

$$C_{b_k}(t)=\sum_{i_k\in\mathcal{I}_k}\theta_{i_k}^{b_k}\left(\rho_{i_k}^{b_k}(t)\right)R_{i_k}(t), \tag{7}$$

where

$$\theta_{i_k}^{b_k}\left(\rho_{i_k}^{b_k}(t)\right)=\begin{cases}1, & \text{if } \rho_{i_k}^{b_k}(t)>0;\\ 0, & \text{otherwise,}\end{cases} \tag{8}$$

is defined as a BS-MU association indicator function.

III. Problem Formulation

With the aforementioned concerns for a HSDRAN, we are particularly interested in investigating how much total BS transmit power is needed in the long term to maintain the queue stability for all MUs, while taking into account the limited fronthaul capacity constraint. Formally, we formulate the following optimization problem:

$$\min_{\{\boldsymbol{\rho}(t):t\in\mathbb{N}^+\}}\sum_{k\in\mathcal{K}}\sum_{b_k\in\mathcal{B}_k}\overline{P}_{b_k} \tag{9a}$$
$$\text{s.t.}\quad \forall k\in\mathcal{K},$$
$$\overline{A}_{i_k}\le\delta\overline{R}_{i_k},\forall i_k\in\mathcal{I}_k; \tag{9b}$$
$$\sum_{b_k\in\mathcal{B}_k}C_{b_k}(t)\le C_{\max},\forall t; \tag{9c}$$
$$P_{b_k}(t)\le P_{\max},\forall b_k\in\mathcal{B}_k,\forall t; \tag{9d}$$
$$\rho_{i_k}^{b_k}(t)\ge 0,\forall b_k\in\mathcal{B}_k,\forall i_k\in\mathcal{I}_k,\forall t. \tag{9e}$$

In (9), $\boldsymbol{\rho}(t)=\{\rho_{i_k}^{b_k}:k\in\mathcal{K},b_k\in\mathcal{B}_k,i_k\in\mathcal{I}_k\}$ is the power adaptation profile for all MUs in the network at time slot t, $\overline{P}_{b_k}=\lim_{\tau\to\infty}(1/\tau)\sum_{t=1}^{\tau}\mathsf{E}[P_{b_k}(t)]$, $\overline{A}_{i_k}=\limsup_{\tau\to\infty}(1/\tau)\sum_{t=1}^{\tau}\mathsf{E}[A_{i_k}(t)]$, $\overline{R}_{i_k}=\liminf_{\tau\to\infty}(1/\tau)\sum_{t=1}^{\tau}\mathsf{E}[R_{i_k}(t)]$, and $C_{\max}$ and $P_{\max}$ are the capacity limit for the fronthaul links and the maximum transit power for the BSs. We refer to $\lambda_{i_k}=\overline{A}_{i_k}$ as the mean data arrival rate of MU i_k. Note that (9b) is the condition of maintaining queue stability for all MUs in the network.

[1]In this paper, the local controllers are assumed to employ a compression-before-precoding strategy [6], namely, the local controllers directly compress the precoding vectors and use the fronthaul links to send the compressed vectors along with the MUs' data messages to the BSs. We further assume that at each time slot, the fronthaul capacity of a BS is mainly consumed by sending the MUs' messages, and the quantization of precoding vectors for BSs is perfect.

Previous approaches usually solve the problem in (9) based on the Markov decision process framework [12] with complete network statistics [14] and suffer from the curse of dimensionality [7]. The introduced constraints of queue stability (9b) and limited fronthaul capacity (9c) make the problem solving even more challenging. Inspired by the penalty-plus-drift method [13], we transform the off-line problem in (9) into an on-line optimization problem by minimizing the bound of one-step Lyapunov drift. Let us define the Lyapunov function as

$$L(t)=\frac{1}{2}\sum_{k\in\mathcal{K}}\sum_{i_k\in\mathcal{I}_k}\left(Q_{i_k}(t)\right)^2, \tag{10}$$

and the one-step conditional Lyapunov drift as

$$\Delta(t)=\mathsf{E}[L(t+1)-L(t)|\mathbf{Q}(t)], \tag{11}$$

where $\mathbf{Q}(t)=\{Q_{i_k}(t):k\in\mathcal{K},i_k\in\mathcal{I}_k\}$ is the global queue state information at each time slot t. Moreover, the drift-plus-penalty function can be defined by

$$\Delta_V(t)=\Delta(t)+V\mathsf{E}\left[\sum_{k\in\mathcal{K}}\sum_{b_k\in\mathcal{B}_k}P_{b_k}(t)\middle|\mathbf{Q}(t)\right], \tag{12}$$

where V is a non-negative weight that trades off the performance. Suppose that for each MU, $A_{i_k}(t)$ and $R_{i_k}(t)$ are upper bounded by $A_{\max}$ and $R_{\max}$, respectively. By squaring both sides of the queue evolution equation (6) and taking the expectation with respect to $\mathbf{Q}(t)$, we deduce

$$\Delta_V(t)\le B+V\mathsf{E}\left[\sum_{k\in\mathcal{K}}\sum_{b_k\in\mathcal{B}_k}P_{b_k}(t)\middle|\mathbf{Q}(t)\right]-\sum_{k\in\mathcal{K}}\sum_{i_k\in\mathcal{I}_k}Q_{i_k}(t)\mathsf{E}\left[\delta R_{i_k}(t)-A_{i_k}(t)|\mathbf{Q}(t)\right], \tag{13}$$

where $B=(1/2)\sum_{k\in K}\sum_{i_k\in\mathcal{I}_k}((R_{\max}\delta)^2+(A_{\max})^2)$. Accordingly, the original transmit power minimization problem in (9) is transformed into solving the following optimization problem at each time slot t,

$$\min_{\boldsymbol{\rho}(t)}\ V\sum_{k\in\mathcal{K}}\sum_{b_k\in\mathcal{B}_k}P_{b_k}(t)-\sum_{k\in\mathcal{K}}\sum_{i_k\in\mathcal{I}_k}\delta Q_{i_k}(t)R_{i_k}(t)$$
$$\text{s.t. constraints (9c), (9d) and (9e)},\forall k\in\mathcal{K}. \tag{14}$$

The key challenge involved in solving above problem arises from the limited fronthaul capacity constraint (9c) for each

cluster. In next section, we aim at designing efficient algorithms for (14) since the convergence speed is a main concern for the per-time slot optimization.

IV. Low-Complexity Algorithm

It can be noticed that the formulated per-time slot optimization problem in (14) is with a non-convex structure due to the introduction of a BS-MU association indicator variable (8), and is considerably different from the problem studied in [15] where the BS-MU association and the radio resource allocation are independent. Without the fronthaul capacity constraint (9c), the optimal solution can be efficiently obtained [16], [17]. For (14), a brute force approach can be applied to achieve the optimal power adaptation. Yet such a method is computationally exhaustive. This motivates us to develop a low-complexity algorithm, which is the focus of this section. In brief, our idea is to first adapt the transmit powers of BSs for a specific set of active BS-MU associations, where all MUs are initially assumed to be served by all BSs in the same cluster (i.e., $\theta_{i_k}^{b_k} = 1$, $\forall b_k \in \mathcal{B}_k$, $\forall i_k \in \mathcal{I}_k$). If the fronthaul capacity constraint (9c) is violated, some of active BS-MU associations are then removed. This process iterates until convergence. In the following, we drop the time slot index t to avoid redundancy.

Specifically, let $\mathcal{B}_{i_k}^{(\ell)} \subseteq \mathcal{B}_k$ be the set of BSs that serve a MU $i_k \in \mathcal{I}_k$ in cluster $k \in \mathcal{K}$ at the ℓ-th iteration with $\ell \in \mathbb{N}_+$ and $\mathcal{B}_{i_k}^{(1)} = \mathcal{B}_k$. Whilst to decouple the power adaptation from the BS-MU association, we make another fundamental assumption that the transmit powers allocated to MU i_k by BSs from $\mathcal{B}_{i_k}^{(\ell)}$ are identical, i.e., $\rho_{i_k}^{b_k} = \rho_{i_k} \in \mathbb{R}_+$, $\forall b_k \in \mathcal{B}_{i_k}^{(\ell)}$. We then solve

$$\min_{\boldsymbol{\rho}} \; V \sum_{k\in\mathcal{K}} \sum_{b_k\in\mathcal{B}_k} P_{b_k} - \sum_{k\in\mathcal{K}} \sum_{i_k\in\mathcal{I}_k} \delta Q_{i_k}(t) R_{i_k} \tag{15a}$$

$$\text{s.t. } \forall k \in \mathcal{K},$$

$$\rho_{i_k}^{b_k} = 0, \quad \forall b_k \notin \mathcal{B}_{i_k}^{(\ell)}, \forall i_k \in \mathcal{I}_k; \tag{15b}$$

$$\rho_{i_k}^{b_k} = \rho_{i_k}, \forall b_k \in \mathcal{B}_{i_k}^{(\ell)}, \forall i_k \in \mathcal{I}_k; \tag{15c}$$

constraints (9d) and (9e),

through the dual decomposition method for each iteration. The fronthaul capacity constraint (9c) can be rewritten as

$$\sum_{b_k\in\mathcal{B}_k} \sum_{i_k\in\mathcal{I}_{b_k}^{(\ell)}} R_{i_k} \le C_{\max}, \tag{16}$$

for cluster $k \in \mathcal{K}$, where $\mathcal{I}_{b_k}^{(\ell)}$ denotes the set of MUs that are served by BS b_k at ℓ-th iteration. Towards this end, we define the Lagrangian function of the primal problem (15) as

$$J(\boldsymbol{\rho};\boldsymbol{\alpha}) = V \sum_{k\in\mathcal{K}} \sum_{b_k\in\mathcal{B}_k} P_{b_k} - \sum_{k\in\mathcal{K}} \sum_{i_k\in\mathcal{I}_k} \delta Q_{i_k} R_{i_k} + \sum_{k\in\mathcal{K}} \sum_{b_k\in\mathcal{B}_k} \alpha_{b_k} \left(P_{b_k} - P_{\max}\right), \tag{17}$$

where $\boldsymbol{\alpha} = \{\alpha_{b_k} : k \in \mathcal{K}, b_k \in \mathcal{B}_k\}$ is a set of the non-negative Lagrangian multipliers (LMs) associated with the maximum transmit power constraint (9d) and the boundary constraint (9e) is absorbed into the Karush-Kuhn-Tucker (KKT) conditions. The dual problem of the primal problem in (15) is given by

$$\max_{\boldsymbol{\alpha}} G(\boldsymbol{\alpha}), \tag{18}$$

where

$$G(\boldsymbol{\alpha}) = \min_{\boldsymbol{\rho}} J(\boldsymbol{\rho};\boldsymbol{\alpha}). \tag{19}$$

Solving (18) directly leads to the optimal solution to (15).

For a given $\boldsymbol{\alpha}$, the KKT conditions are necessary and sufficient for (19). By rearranging (17), we have an aggregated version of the Lagrangian function as

$$J(\boldsymbol{\rho};\boldsymbol{\alpha}) = \sum_{k\in\mathcal{K}} \left(\sum_{i_k\in\mathcal{I}_k} J_{i_k}(\boldsymbol{\rho};\boldsymbol{\alpha}_k) + \sum_{b_k\in\mathcal{B}_k} \alpha_{b_k} P_{\max} \right), \tag{20}$$

where $\boldsymbol{\alpha}_k = \{\alpha_{b_k} : b_k \in \mathcal{B}_k\}$ and

$$J_{i_k}(\boldsymbol{\rho};\boldsymbol{\alpha}_k) = \sum_{b_k\in\mathcal{B}_{i_k}^{(\ell)}} \left(V + \alpha_{b_k}\right) \rho_{i_k} - \delta Q_{i_k} R_{i_k}, \tag{21}$$

which incorporates constraints (15b) and (15c). This enables us to solve (15) in a distributed way, hence alleviating the information exchange overheads between high-level controller and local controllers. We formulate the power adaptation among the clusters as a non-cooperative game, in which the local controllers are the players, the action of each local controller $k \in \mathcal{K}$ is the BS power adaptations $\boldsymbol{\rho}_k = \{\rho_{i_k}^{b_k} : b_k \in \mathcal{B}_k, i_k \in \mathcal{I}_k\}$ and the payoff function is defined as $J_k(\boldsymbol{\rho}_k, \boldsymbol{\rho}_{-k}; \boldsymbol{\alpha}_k) = \sum_{i_k\in\mathcal{I}_k} J_{i_k}(\boldsymbol{\rho}_k, \boldsymbol{\rho}_{-k}; \boldsymbol{\alpha}_k) + \sum_{b_k\in\mathcal{B}_k} \alpha_{b_k} P_{\max}$, where $\boldsymbol{\rho}_{-k} = \{\boldsymbol{\rho}_{k'}' : k' \in \mathcal{K} \setminus \{k\}\}$. The non-cooperative power adaptation game can be characterized by

$$\min_{\boldsymbol{\rho}_k} J_k(\boldsymbol{\rho}_k, \boldsymbol{\rho}_{-k}; \boldsymbol{\alpha}_k), \forall k \in \mathcal{K}. \tag{22}$$

In cluster k, the power adapted by a BS $b_k \in \mathcal{B}_{i_k}^{(\ell)}$ for serving a MU $i_k \in \mathcal{I}_k$ can be expressed as a function of $\boldsymbol{\rho}_{-k}$ and $\boldsymbol{\alpha}_k$, namely,

$$\rho_{i_k} = \varphi_{i_k}(\boldsymbol{\rho}_{-k}; \boldsymbol{\alpha}_k) = \left[\frac{\delta W Q_{i_k}}{\ln(2) \sum_{b_k\in\mathcal{B}_{i_k}^{(\ell)}} (V + \alpha_{b_k})} - \frac{1}{\Omega_{i_k}(\boldsymbol{\rho}_{-k})} \right]^+, \tag{23}$$

where $\Omega_{i_k}(\boldsymbol{\rho}_{-k}) = |\sum_{b_k\in\mathcal{B}_{i_k}^{(\ell)}} (\mathbf{h}_{i_k}^{b_k})^{\mathsf{H}} \mathbf{u}_{i_k}^{b_k}|^2 / (M_{i_k} + \sigma^2)$. The Nash equilibrium solution to (22) can be expressed as $\rho_{i_k}^* = \varphi_{i_k}(\boldsymbol{\rho}_{-k}^*; \boldsymbol{\alpha}_k)$, $\forall k \in \mathcal{K}$, $\forall i_k \in I_k$. The optimal LMs can be found using an incremental-update based sub-gradient method [18], which in general converges slowly. From (23), the power adaptation is a strictly decreasing function of the LMs. The bisection approach [10] can be accordingly explored to find the LMs that ensure (9d). The power adaptation at each ℓ-th iteration is summarized in Algorithm 1.

The outcomes from Algorithm 1 do not take into account the fronthaul capacity constraint (16). If the power adaptation at the ℓ-th iteration satisfies the fronthaul capacity constraint (16) for all clusters, the process terminates. Otherwise, if the

Algorithm 1 Power adaptation at each local controller $k \in \mathcal{K}$ during each ℓ-th iteration

1: **initialize** $\boldsymbol{\alpha}_{k,\min} = \{\alpha_{b_k,\min} : b_k \in \mathcal{B}_k\}$ and $\boldsymbol{\alpha}_{k,\max} = \{\alpha_{b_k,\max} : b_k \in \mathcal{B}_k\}$ for local controller k.
2: **repeat**
3: Calculate the transmit power ρ_{i_k} allocated by BSs in $\mathcal{B}_{i_k}^{(\ell)}$ to each MU $i_k \in \mathcal{I}_k$ according to (23).
4: Calculate the total transmit power P_{b_k} of each BS $b_k \in \mathcal{B}_k$ across MUs in $\mathcal{I}_{b_k}^{(\ell)}$ according to (5).
5: For a BS $b_k \in \mathcal{B}_k$, if $P_{b_k} \leq P_{\max}$, set $\alpha_{b_k} = \alpha_{b_k,\min}$; otherwise, update α_{b_k} using the bisection approach.
6: **until** Convergence

fronthaul capacity limit in a cluster $k \in \mathcal{K}$ is exceeded, we need to remove the fronthaul for one BS-MU association, say $(b_k^{(\ell)}, i_k^{(\ell)})$, and solve (15) at next iteration $\ell+1$ with updated

$$\mathcal{B}_{i_k}^{(\ell+1)} = \begin{cases} \mathcal{B}_{i_k}^{(\ell)} \setminus \left\{b_k^{(\ell)}\right\}, & \text{if } i_k = i_k^{(\ell)}; \\ \mathcal{B}_{i_k}^{(\ell)}, & \text{otherwise.} \end{cases} \tag{24}$$

The problem left is how to select a BS-MU association to be shut down. In this paper, based on the power adaptation results at ℓ-th iteration, we propose a criterion in the following

$$\left(b_k^{(\ell)}, i_k^{(\ell)}\right) = \arg\min_{(b_k,i_k)\in\mathcal{B}_{i_k}^{(\ell)}\times\mathcal{I}_k} \left\{\tilde{J}_{i_k,-b_k}(\boldsymbol{\rho}_{i_k};\boldsymbol{\alpha}_k) - J_{i_k}(\boldsymbol{\rho}_{i_k};\boldsymbol{\alpha}_k)\right\}, \tag{25}$$

where

$$\tilde{J}_{i_k,-b_k}(\boldsymbol{\rho}_{i_k};\boldsymbol{\alpha}_k) = \sum_{b_k'\in\mathcal{B}_{i_k}^{(\ell)}\setminus\{b_k\}} \left(V + \alpha_{b_k'}\right)\rho_{i_k} - \delta Q_{i_k}\tilde{R}_{i_k,-b_k}, \tag{26}$$

with

$$\tilde{R}_{i_k,-b_k} = W\log_2\left(1 + \frac{\rho_{i_k}\left|\sum_{b_k'\in\mathcal{B}_{i_k}^{(\ell)}\setminus\{b_k\}}\left(\mathbf{h}_{i_k}^{b_k'}\right)^{\mathsf{H}}\mathbf{u}_{i_k}^{b_k'}\right|^2}{M_{i_k}+\sigma^2}\right). \tag{27}$$

Intuitively, the proposed criterion is to choose the BS-MU association that the removal results in the least increase of the Lagrangian function. In Algorithm 2, we present the proposed solution for fronthaul capacity constrained power adaptation at each time slot t.

V. Numerical Results

In this section, we carry out Matlab-based simulation experiments to evaluate the proposed work. During simulations, the considered HSDRAN is assumed to consist of 4 clusters which cover an $2\mathrm{Km}\times 2\mathrm{Km}$ square area. Within each cluster, the MUs are uniformly distributed. The channel gains for links between a MU and the antennas of a BS is randomly generated at each time slot according to the channel model used in [19].

Algorithm 2 Fronthaul capacity constrained power adaptation at each time slot t

1: **initialize** $\mathcal{B}_{i_k}^{(1)} = \mathcal{B}_k$ for all $i_k \in \mathcal{I}_k$ in each cluster $k \in \mathcal{K}$.
2: **repeat**
3: Solve Problem (15) using Algorithm 1, obtain power adaptation $\boldsymbol{\rho}_{i_k}$ for each MU $i_k \in \mathcal{I}_k$.
4: If fronthaul capacity constraint (16) does not hold, update $\mathcal{B}_{i_k}^{(\ell+1)}$ according to (24) and set $\ell = \ell + 1$. Otherwise, terminate the algorithm.
5: **until** Convergence
6: Set the optimal power adaptation at time t to be $\boldsymbol{\rho}^*(t) = \boldsymbol{\rho}$.

TABLE I
Simulation Parameters.

Parameter	Value
Maximum transmit power $P_{\max}$	30 dBm
Number of clusters K	4
Number of BSs per cluster $\lvert\mathcal{B}_k\rvert$	3, $\forall k \in \mathcal{K}$
Number of antennas per BS N	2
Number of MUs per cluster $\lvert\mathcal{I}_k\rvert$	3, $\forall k \in \mathcal{K}$
Spectral bandwidth W	10 MHz
Noise power spectral density σ^2	-169 dBm/Hz
Time slot duration δ	10^{-3} second
Shape factor ζ	1.2
Mode υ	2×10^3 bits
Cutoff threshold q	10^4 bits

We use the Pareto distribution to simulate the traffic arrivals at the MU side [20]. More specifically, the probability that a data bits arrive during a time slot can be expressed as

$$f(a) = \begin{cases} \frac{\zeta\upsilon^\zeta}{a^{\zeta+1}}, & \text{if } \upsilon \leq a < q; \\ \nu, & \text{if } a \geq q, \end{cases} \tag{28}$$

where ζ is the shape factor, υ is the mode, q is the cutoff threshold, and ν is calculated as

$$\nu = \left(\frac{\upsilon}{q}\right)^\zeta, \tag{29}$$

for $\zeta > 1$. The parameter values used in the simulations are listed in Table I.

In the first experiment, we fix the fronthaul capacity limit for all clusters to be $C_{\max} = 300$ Mbits and the value of V as 10^9. In this case, Fig. 2 shows that both the queue length of MUs and the sum transmit power of BSs in one of the clusters dynamically converge. The second experiment examines the proposed algorithm under different values of $C_{\max}$ and V. From the upper subplot in Fig. 3, we can see that as $C_{\max}$ increases, the average queue length achieved by MUs decreases. The reason behind this can be easily understood. As the fronthaul capacity limit increases, the MUs can be associated with more BSs (or allocated more transmit power from BSs), and thus realize higher data rate. We also find from the subplot that with a larger V value, the MUs achieve longer average queue length since the transmit power minimization is of higher priority than the queue stability, which is confirmed

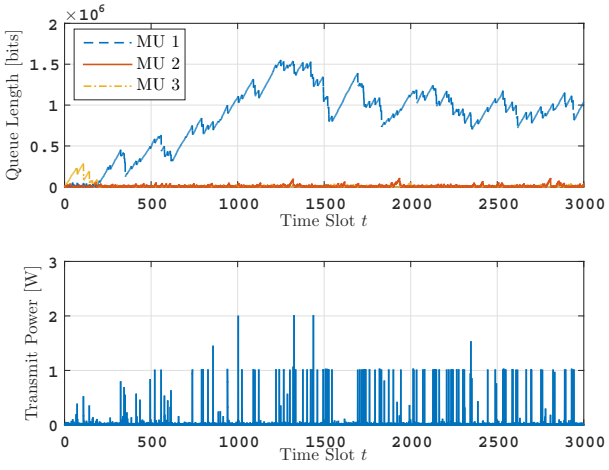

Fig. 2. Queue length of MUs and sum transmit power of BSs in a cluster k versus time slot.

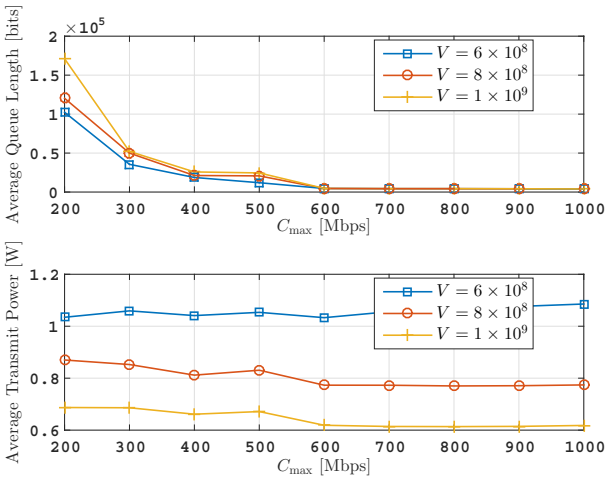

Fig. 3. Average queue length of MUs and average sum transmit power of BSs in the network versus fronthaul capacity.

by the lower subplot in Fig. 3 as well. However, from the lower subplot in Fig. 3, with a sufficiently small V value, the average sum transmit power increases as $C_{\max}$ increases. This is due to the fact that a smaller V value puts more importance of stabilizing the MU data queues.

VI. Conclusions

In this paper, we investigate a hierarchical architecture design of the software-define RANs and the problem of long-term downlink power adaptation in an HDRAN. The introduced queue stability constraint for MUs and limited fronthaul capacity constraint for each cluster make the problem solving extremely challenging. We show that the drift-plus-penalty method can achieve the queue stability for all MUs and the sum transmit power minimization of BSs across different clusters and thereby transforms the off-line stochastic transmit power adaptation into an on-line per time slot optimization problem. To solve the non-convex per time slot optimization and avoid extensive information exchange between the high-level controller and the local controllers, we propose a low-complexity algorithm by formulating a non-cooperative power adaptation game among the local controllers. Preliminary simulation experiments illustrate that our proposed algorithm can achieve optimal required transmit power while the MU queue stability being ensured.

References

[1] H. Taoka, "Views on 5g," Tech. Rep., Dusseldorf, Germany, Oct. 2011.

[2] J. Wu, S. Rangan, and H. Zhang, *Green Communications: Theoretical Fundamentals, Algorithms and Applications*. Boca Raton, FL, USA: CRC Press, Sep. 2012.

[3] A. Gudipati, D. Perry, L. E. Li, and S. Katti, "SoftRAN: Software defined radio access network," in *ACM SIGCOMM HotSDN Workshop*, Hong Kong, China, Aug. 2013.

[4] T. Chen, H. Zhang, X. Chen, and O. Tirkkonen, "SoftMobile: Control evolution for future heterogeneous mobile networks," *IEEE Wireless Commun.*, vol. 21, no. 6, pp. 70–78, Dec. 2014.

[5] L. E. Li, Z. M. Mao, and J. Rexford, "Toward software-defined cellular networks," in *EWSDN*, Darmstadt, Germany, Oct. 2012.

[6] J. Kang, O. Simeone, J. Kang, and S. Shamai, "Fronthaul compression and precoding design for C-RANs over ergodic fading channels," *IEEE Tran. Veh. Technol.*, vol. 65, no. 7, pp. 5022–5032, Jul. 2016.

[7] Y. Cui, Q. Huang, and V. K. N. Lau, "Queue-aware dynamic clustering and power allocation for network MIMO systems via distributed stochastic learning," *IEEE Tran. Signal Processing*, vol. 59, no. 3, pp. 1229–1238, Mar. 2011.

[8] V. N. Ha, L. B. Le, and N.-D. Đào, "Coordinated multipoint (CoMP) transmission design for cloud-RANs with limited fronthaul capacity constraints," *IEEE Tran. Veh. Technol.*, vol. 65, no. 9, pp. 7432–7447, Sep. 2016.

[9] L. Liang and R. Zhang, "Macrocell-queue-stabilization-based power control of femtocell networks," in *Proc. IEEE ICASSP*, Shanghai, China, Mar. 2016.

[10] H. Wang and Z. Ding, "Downlink SINR balancing in C-RAN under limited fronthaul capacity," *IEEE Trans. Wireless Commun.*, vol. 13, no. 9, pp. 5223–5236, Sep. 2014.

[11] M. J. Neely, "Optimal peer-to-peer scheduling for mobile wireless networks with redundantly distributed data," *IEEE Trans. Mobile Comput.*, vol. 13, no. 9, pp. 2086–2099, Sep. 2014.

[12] E. Altman, *Constrained Markov Decision Processes*. London, UK: Chapman & Hall/CRC, 1999.

[13] M. J. Neely, "Stochastic network optimization with application to communication and queueing systems," *Synthesis Lectures on Communication Networks: ch. 1*, San Rafael, CA: Morgan & Claypool, 2010.

[14] L. Saker, S.-E. Elayoubi, R. Combes, and T. Chahed, "Optimal control of wake up mechanisms of femtocells in heterogeneous networks," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 664–672, Apr. 2012.

[15] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "User association for load balancing in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2706–2716, Jun. 2013.

[16] X. Zhang, Y. Sun, X. Chen, S. Zhou, J. Wang, and N. B. Shroff, "Distributed power allocation for coordinated multipoint transmissions in distributed antenna systems," *IEEE Trans. Wireless Commun.*, vol. 12, no. 5, pp. 2281–2291, May 2013.

[17] J. Li, E. Björnson, T. Svensson, T. Eriksson, and M. Debbah, "Joint precoding and load balancing optimization for energy-efficient heterogenous networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 10, pp. 5810–5822, Oct. 2015.

[18] M. Hajiaghayi, M. Dong, and B. Liang, "Optimal channel assignment and power allocation for dual-hop multi-channel multi-user relaying," in *IEEE INFOCOM*, Shanghai, China, Apr. 2011.

[19] Y. Shi, J. Zhang, B. O'Donoghue, and K. B. Letaief, "Large-scale convex optimization for dense wireless cooperative networks," *IEEE Tran. Signal Processing*, vol. 63, no. 18, pp. 4729–4743, Sep. 2015.

[20] N. Salodkar, A. Karandikar, and V. S. Borkar, "A stable online algorithm for energy-efficient multiuser scheduling," *IEEE Trans. Mobile Comput.*, vol. 9, no. 10, pp. 1391–1406, Oct. 2010.