

Energy-Efficient Resource Allocation Optimization for Multimedia Heterogeneous Cloud Radio Access Networks

Mugen Peng, *Senior Member, IEEE*, Yuling Yu, Hongyu Xiang, and H. Vincent Poor, *Fellow, IEEE*

Abstract—The heterogeneous cloud radio access network (H-CRAN) is a promising paradigm which incorporates the cloud computing into heterogeneous networks (HetNets), thereby taking full advantage of cloud radio access networks (C-RANs) and HetNets. Characterizing the cooperative beamforming with fronthaul capacity and queue stability constraints is critical for multimedia applications to improving energy efficiency (EE) in H-CRANs. An energy-efficient optimization objective function with individual fronthaul capacity and inter-tier interference constraints is presented in this paper for queue-aware multimedia H-CRANs. To solve this non-convex objective function, a stochastic optimization problem is reformulated by introducing the general Lyapunov optimization framework. Under the Lyapunov framework, this optimization problem is equivalent to an optimal network-wide cooperative beamformer design algorithm with instantaneous power, average power and inter-tier interference constraints, which can be regarded as the weighted sum EE maximization problem and solved by a generalized weighted minimum mean square error approach. The mathematical analysis and simulation results demonstrate that a tradeoff between EE and queuing delay can be achieved, and this tradeoff strictly depends on the fronthaul constraint.

Index Terms—Heterogeneous cloud radio access networks, multimedia traffic, queue-aware, Lyapunov optimization.

I. INTRODUCTION

With the explosive growth of mobile multimedia traffic demand and number of mobile devices, the next-generation wireless networks face significant challenges in improving system capacity and guaranteeing users' quality of service (QoS). Cloud radio access networks (C-RANs) present a promising approach to these challenges by curtailing both capital and operating expenditures for providing mobile multimedia applications, while providing high energy-efficiency and capacity [1] [2]. In C-RANs, the traditional base station (BS) is decoupled into the distributed remote radio heads (RRHs) and the baseband unit (BBU). Antennas are equipped with RRHs to transmit/receive radio frequency (RF) signals, and BBUs are clustered as a BBU pool in a centralized location with aggregating all BS computational resources, which provides large-scale processing and management functions for the signals transmitted/received from RRHs. With this architecture,

mobile operators can easily expand and upgrade the network by deploying additional RRHs, and thus the corresponding operational costs can be greatly reduced.

The heterogeneous cloud radio access network (H-CRAN) is regarded as a new paradigm to meet performance requirements of the fifth generation (5G) cellular system for mobile multimedia applications by incorporating cloud computing into heterogeneous networks (HetNets) [3] [4], in which the control and user planes are decoupled. The existing macro base station (MBS) that has been deployed in traditional cellular networks is used to alleviate capacity constraints over the fronthaul and provide seamless coverage with QoS guarantees for users. **In particular, burst multimedia traffic and real-time multimedia traffic with low-bit transmit rate can be efficiently served by the MBS.** For control signaling and system data broadcasting at MBSs, it alleviates the capacity and time delay constraints in the fronthaul links between RRHs and the BBU pool, and allows RRHs to use sleep mode efficiently to decrease energy consumption. **RRHs are preferred to provide both real-time and non-real time multimedia applications with high speed data rates, such as real-time interactive high quality video, delay-tolerant web browsing, non-real time video or massive file download, etc.** With the help of MBSs, RRHs can be used to provide only the high-capacity service and are transparent to the served users. Note that the radio signal processing for all RRHs is executed in the BBU pool, while for the MBS is implemented locally. The inter-tier interference between the BBU pool and the MBS can be mitigated by the distributed coordinated multi-point (CoMP) transmission and reception technique. Comparing with C-RANs and HetNets, H-CRANs have been demonstrated to significant performance gains though advanced collaborative signal processing and cooperative radio resource allocation are still challenging [3].

Intuitively, cloud computing in the BBU pool based on large-scale cooperative signal processing can suppress intra-tier interference and achieve significant cooperative gains in H-CRANs. **The inter-tier interference to RRHs from the MBS equipped with multiple antennas can be suppressed by coordinated scheduling or cooperative multiple-input multiple-output (MIMO) techniques, which substantially improves the spectral efficiency (SE).** For instance, inter-tier interference can be suppressed by using zero-forcing, which results from the **aggressive spatial multiplexing [5].** Such characteristics in H-CRANs bring challenges to optimize the overall SE or energy efficiency (EE) because too many factors and challenges must be jointly considered, such as collaborative signal processing

Mugen Peng, Yuling Yu, and Hongyu Xiang are with the Key Laboratory of Universal Wireless Communications for Ministry of Education, Beijing University of Posts and Telecommunications, Beijing, China (e-mail: pmg@bupt.edu.cn, aliceyu1215@gmail.com, xianghongyu88@gmail.com).

H. V. Poor is with the School of Engineering and Applied Science, Princeton University, Princeton, NJ, USA (e-mail: poor@princeton.edu).

to suppress intra-tier and inter-tier interference in the physical (PHY) layer, and cooperative radio resource allocation and queue-aware packet scheduling in the medium access control (MAC) and upper layers. In addition, capacity constraints of fronthaul and backhaul links must be considered as well.

A. Related Work

Much attention has been paid to resource allocation in C-RANs recently. In [6], to minimize the network power consumption, a greedy RRH on/off selection algorithm has been proposed to maximize the reduction in the network power consumption at each step. In [7], an antenna selection scheme that jointly optimizes the antenna selection, regularization factor and power allocation has been presented to maximize the averaged weighted sum-rate in large-scale C-RAN downlink systems. The joint optimization of MIMO and discontinuous transmission (DTX) with practical implementation constraints has been investigated in [8] to improve EE performance. Meanwhile, [9] has proposed a joint cell association and beamformer design algorithm for downlink and uplink C-RANs. Clearly, these characteristics and achievements to improve SE and EE performance of C-RANs should be further enhanced in H-CRANs, in which the cell association with RRH/MBS and the inter-tier interference should be additionally considered.

Meanwhile, a number of studies have considered the SE and EE optimization of HetNets, in which radio resource allocation with the consideration of inter-tier interference is often the primary focus. In [10], an EE optimization problem with statistical quality of service (QoS) constraints in orthogonal frequency-division multiple access (OFDMA) systems has been analyzed, and a subchannel grouping scheme to obtain a closed form solution has been presented, which is simplified to a multi-target single-channel optimization problem by using the channel-matrix singular value decomposition method. In [11], to improve EE in heterogeneous cognitive femtocell networks, a spectrum sharing and resource allocation scheme has been formulated as a Stackelberg game, and a gradient based iterative algorithm has been proposed to achieve the Stackelberg equilibrium solution. In [12], an energy-efficient partial spectrum reuse (PSR) scheme has been proposed. Since the optimal PSR factor, defined as the portion of spectrum reused by micro cells in two-tier heterogeneous networks, is not in an explicit form generally, a closed-form limit of the optimal PSR factor has been derived as the ratio of the user rate requirement to the entire system spectrum bandwidth. Numerical results showed that adopting PSR can reduce the network energy consumption by up to 50% when the transmit power of MBSs is 10dB higher than that of low power nodes (LPNs). An energy efficient precoding for coordinated multi-point transmission under constraints of individual data rate requirements from each user, maximal transmit power of each base station (BS), and zero-forcing (ZF) has been obtained by introducing the subspace decomposition method in [13], where the performance gain of the proposed ZF precoder has been verified by comparing with several existing optimal linear precoders.

The aforementioned works typically assume that all users are time-delay insensitive and neglect special QoS require-

ments for delay sensitive users, which may suffer serious performance deterioration caused by the large service delay. To minimize the user's queuing time and achieve a degree of fairness, the authors of [14] designed an efficient blind scheduling policy that performs well across magnitudes of fairness, simplicity, and asymptotic optimality for a relatively general mobile media cloud. In [15], a joint power and rate control algorithm with average delay constraints has been proposed and solved by a game-theoretic model. To study energy efficiency-delay tradeoffs in multiple-access networks, a game-theoretic approach is proposed in [16], where each user seeks to choose a transmit power that maximizes its own utility while satisfying its delay requirements. To deal with the co-channel interference problem and the individual statistical delay QoS guarantee problem, the cross-layer optimization of a two-tier underlay HetNet has been studied in [17] using large deviation theory, in which the cross-layer optimization problem is transformed into a long term weighted sum effective capacity maximization problem. In addition, to further guarantee the service fairness between different delay-tolerant users, the queue backlogs maintained at the transmitters for each user have to be considered in the design of radio resource optimization schemes. The authors of [18] have investigated the delay-optimal policy in a two-user multiple access channel, where the delay minimization problem is formulated as a Markov decision process (MDP) and the optimal policy traded a portion of the sum-rate for balancing the queue lengths to minimize the average delay. The delay-optimal power and sub-carrier allocation problem for OFDMA systems was modeled as a K -dimensional infinite horizon average reward MDP with the control actions based on channel state information (CSI) and joint queue state information (QSI) in [19]. Furthermore, a cache-enabled cross-layer opportunistic cooperative MIMO framework for wireless video streaming is proposed in [20]. By equipping the relay with a cache to buffer the video streams, the cache control policy adaptive to the popularity of the video files could provide more cooperative opportunities, while the power control policy adaptive to QSI and CSI is determined by solving the approximated MDP approach using the continuous time Bellman equations to maintain the QoS metrics of playback interruption probability and the buffer overflow probability.

However, the number of queues in realistic systems is not often sufficiently large, which causes the issue of curse of dimensionality with the MDP approach due to the exponential growth of the cardinality of the system state space. In addition, it is difficult to obtain a distributed resource allocation solution with MDP since the potential function is not decomposable. To achieve a desired tradeoff between network throughput and queuing delay, [21] and [22] proposed distributed resource allocation and user scheduling solutions in Long Term Evolution-Advanced (LTE-A) relay networks and wireless multihop networks, respectively, both of which utilize Lyapunov optimization to stabilize the queues of networks when optimizing performance metrics. Lyapunov optimization is a useful tool for handling queue-aware radio resource allocation problems with a good balance between performance and implementation complexity. For the Lyapunov optimization

theory, the concepts of Lyapunov function and Lyapunov drift are introduced, and the performance metrics of the network can be optimized while stabilizing queues of the network by greedily minimizing the drift-plus-penalty. With the Lyapunov optimization approach, the problem of opportunistic cooperation in a cognitive two-tier underlay HetNet has been studied in [23], where the cognitive LPNs handle intelligent access admission, cooperation decision making, and power control to maximize their own throughputs subject to average power constraints. The obtained online control algorithm can stabilize the multimedia traffic queue without requiring any knowledge of the multimedia traffic arrival rates. A two-stage queue-aware cross-layer radio resource allocation algorithm has been proposed in [24] based on minimizing drift-plus-utility, which can be easily applied for HetNets.

Inspired by [21]– [24], the well-developed stability theory of Lyapunov optimization is considered in this paper. Since the optimal radio resource allocation policy can be achieved by minimizing the drift-plus-utility, in which the resulting queuing delay and utility performance are bounded, we focus on maximizing EE under the stable queue with the transmit power, individual fronthaul capacity, and interference constraints in H-CRANs. There are three technical challenges associated with this EE optimization problem in queue-aware H-CRANs:

- **Challenges due to Joint Considerations of EE and Queuing Delay:** Unlike other works focusing only on optimizing SE or EE, which only requires CSI in the PHY layer, the optimization involving EE and average queuing delay is fundamentally challenging since it introduces coupling between the PHY and MAC layers. To take the queuing delay into consideration, the resource allocation policy should be a function of both QSI and CSI, which is a nontrivial problem since the QSI and CSI vary randomly at each time slot and may not have a closed-form expression.
- **Challenges due to Inter-tier Interference and Individual Fronthaul Capacity Constraints:** Unlike traditional C-RANs, the inter-tier interference from MBSs in H-CRANs should be suppressed by advanced collaborative processing techniques, and the inter-tier interference to MUEs should be mitigated to a low level with advanced coordinated scheduling and power control techniques. Unlike the traditional HetNets, intra-tier interference among RRHs in H-CRANs can be suppressed through the centralized BBU pool but with individual fronthaul capacity constraints, and the inter-tier interference often exists between a single powerful MBS and a very large number of RRHs.
- **Challenges due to Minimization of Drift-plus-penalty:** The queues of data flows are coupled due to the mutual inter-tier interference in H-CRANs. Thus, the associated stochastic optimization problem formulated by minimizing the drift-plus-penalty under the Lyapunov optimization framework is complex because resource allocation decisions for different RRHs are affected by each other. With the time-varying nature of wireless environments,

this problem is challenging to solve.

B. Contribution and Organization

With the introduction of H-CRANs, the development of effective radio resource management techniques to optimize EE for non-real time packet service is important. In addition, to satisfy diverse QoS requirements, it is crucial to use cross-layer radio resource management algorithms in H-CRANs, which has been seldom studied to date. In this paper, a weighted EE performance metric is presented, and the corresponding EE optimization problem with inter-tier interference, individual fronthaul capacity, and total transmission power constraints is formulated, in which both cross-layer design and queue-aware congestion control are taken into account. Since this EE optimization problem is a combination of time-averaged variables and instantaneous variables, an optimal network-wide cooperative beamformer design algorithm is proposed based on minimizing the drift-plus-utility under the Lyapunov optimization framework, in which a generalized weighted minimum mean square error (WMMSE) [25] approach is used to optimize the EE performance and guarantee the queue stability. Although the WMMSE approach has been applied in [26] to jointly optimize the user scheduling and beamforming vectors under either dynamic or fixed BS clustering for C-RANs, this previous work does not explicitly take the EE optimization and queue stability into consideration for H-CRANs.

The major contributions of this paper can be summarized as follows.

- An average weighted EE utility function in terms of sum transmit rate and total energy consumption with different weight factors is defined to capture the EE performance in H-CRANs. To maximize this EE performance metric and keep the multimedia traffic queue stability, an EE optimization problem with the instantaneous and average power, individual fronthaul capacity and inter-tier interference constraints is formulated for queue-aware H-CRANs. To solve this non-convex optimization problem, the Lyapunov optimization framework is utilized, under which the optimization problem is transformed into the minimization of the drift-plus-penalty function. Furthermore, this minimization of the drift-plus-penalty function can be reformulated as the optimal network-wide beamformer design problem under transmit power and inter-tier interference constraints.
- A generalized WMMSE approach is proposed to solve the optimal network-wide beamformer design problem. Unlike previous work in [26], whose aim is to solve the weighted sum rate maximization problem with backhaul constraints in C-RANs, this paper applies the generalized WMMSE approach to solve the average weighted EE utility objective function with each RRH's transmit power, individual fronthaul capacity, and inter-tier interference constraints. To quantitatively optimize the tradeoff between the average weighted EE and the queuing delay on demand, a non-negative parameter V is defined, which in turn adaptively affects the solutions of network-wide cooperative beamformer design and power allocation.

- The proposed optimal network-wide beamformer design algorithm can approach an $[\mathcal{O}(1/V), \mathcal{O}(V)]$ tradeoff between the averaged weighted EE performance and queue backlog, which indicates that the average weighted EE performance can be arbitrarily close to the optimum with the gap of $\mathcal{O}(1/V)$ at the expense of incurring an average queue backlog that is $\mathcal{O}(V)$. Simulation results exhibit the EE performance under varied V , and show the influence of the fronthaul capacity constraint on the tradeoff between the average weighted EE performance and average queue backlog.

The rest of this paper is organized as follows. Section II describes the system model and formulates the optimization problem. In Section III, based on the general Lyapunov optimization framework, the formulated problem is transformed into a WMMSE problem which can be solved by an iterative approach. In Section IV, the performance of the proposed network-wide beamformer design algorithm is analyzed. Section V presents the simulation results and Section VI summarizes this paper.

Throughout this paper, the following notation is adopted. **Lower-case bold letters \mathbf{v} denote column vectors, and upper-case bold letters \mathbf{D} denote matrices.** We use \mathbb{C} to denote the complex domain. The complex Gaussian distribution is represented by $\mathcal{CN}(\cdot, \cdot)$, while $\mathbf{Re}\{\cdot\}$ stands for the real part of a scalar. We use $\{x\}^+$ to denote the larger of x and 0. $\mathbb{E}[x]$ is the expectation of the random variable x , and $(\cdot)^H$ denotes the matrix conjugate transpose. $\mathbf{0}_N$ and \mathbf{I}_N are $N \times N$ zero matrix and identity matrix, respectively.

II. SYSTEM MODEL AND PROBLEM FORMULATION

In this section, the considered H-CRAN system scenario and definition of network stability are introduced. Based on the H-CRAN system model and defined queue stability, the EE optimization problem is formulated.

A. System Model

As illustrated in Fig. 1, a downlink H-CRAN system consisting of one MBS and N RRHs is considered. N RRHs are deployed within the same coverage of the single MBS in an underlay manner. The RRHs and MBS are connected to a BBU pool with the fronthaul and backhaul links, respectively. The MBS and each RRH are equipped with L_M and L_R antennas, respectively. Define the set of MBS and RRHs as $\{0, 1, 2, \dots, N\}$, where the index 0 refers to the MBS, which serves K_M single-antenna MBS user equipments (MUEs), and $\mathcal{N} = \{1, 2, \dots, N\}$ denotes the set of RRHs, which cooperatively serve K_R single-antenna RRH user equipments (RUEs) with user-centric clustering. Define the set of RUEs as $\mathcal{K}_R = \{1, 2, \dots, K_R\}$, and the set of MUEs as $\mathcal{K}_M = \{1, 2, \dots, K_M\}$. This H-CRAN system is assumed to operate in the slotted time mode with the unit time slot $t \in \{0, 1, 2, \dots\}$, where the time slot t refers to the interval $[t, t+1)$. Under the assumption that the BBU pool centrally processes all RUEs' signals and distributes each RUE's data to an individually selected cluster of RRHs through fronthaul links, each RUE is cooperatively served by its serving cluster through the joint beamforming technique, and receives an independent data

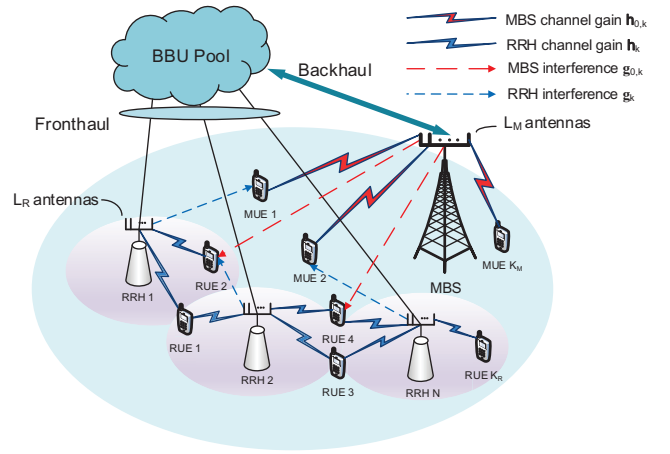


Fig. 1. Downlink H-CRANs with one MBS, N RRHs, K_R RUEs and K_M MUEs.

stream from the RRH at the time slot t . It is assumed that the scalar-valued data stream $s_k(t)$ is temporally white with zero mean and unit variance.

Under centralized large-scale cooperative processing in the BBU pool, the transmit beamformer from RRH n to RUE k in the time slot t is defined as $\mathbf{v}_{n,k}(t) \in \mathbb{C}^{L_R \times 1}$, and the corresponding network-wide beamforming vector for RUE k can be expressed as $\mathbf{v}_k(t) = [\mathbf{v}_{1,k}^H(t), \mathbf{v}_{2,k}^H(t), \dots, \mathbf{v}_{N,k}^H(t)]^H \in \mathbb{C}^{N L_R \times 1}$. Given that $\mathbf{D}_n = \left\{ \underbrace{\mathbf{0}_{L_R}, \dots, \mathbf{0}_{L_R}}_{n-1}, \mathbf{I}_{L_R}, \dots, \mathbf{0}_{L_R} \right\} \in \mathbb{C}^{L_R \times N L_R} (n > 0)$, $\mathbf{v}_{n,k}(t)$ can be represented through $\mathbf{v}_k(t)$, i.e.,

$$\mathbf{v}_{n,k}(t) = \mathbf{D}_n \mathbf{v}_k(t), \quad (1)$$

Note that $\mathbf{v}_k(t)$ can be written as a combination of the transmit power ($\|\mathbf{v}_k(t)\|^2$) and the unit beamformer ($\bar{\mathbf{v}}_k(t) = \frac{\mathbf{v}_k(t)}{\|\mathbf{v}_k(t)\|}$) for simplicity.

In particular, the transmit beamformer from the MBS to MUE k is denoted by $\mathbf{v}_{0,k}(t) \in \mathbb{C}^{L_M \times 1}$. Though all RRHs can potentially serve each scheduled RUE, in fact, each RUE is mainly contributed to by only a small number of adjacent RRHs and the network-wide beamforming vector is often group sparse [26].

With the linear transmit beamforming scheme at the RRHs [27], the received signal at the RUE k , denoted by $y_k(t) \in \mathbb{C}$, consists of the desired signal, the interference signal of other RUEs, and the interference signal of the total K_M MUEs. As a result, $y_k(t) \in \mathbb{C}$ can be written as

$$y_k(t) = \mathbf{h}_k^H(t) \mathbf{v}_k(t) s_k(t) + \sum_{j=1, j \neq k}^{K_R} \mathbf{h}_k^H(t) \mathbf{v}_j(t) s_j(t) + \sum_{i=1}^{K_M} \mathbf{g}_{0,k}^H(t) \mathbf{v}_{0,i}(t) s_i(t) + n_k(t), \quad (2)$$

where $\mathbf{h}_k(t) \in \mathbb{C}^{N L_R \times 1}$ denotes the CSI matrix from all RRHs' transmit antennas to the RUE k , and $\mathbf{g}_{0,k}(t) \in \mathbb{C}^{L_M \times 1}$ denotes the CSI matrix from the MBS's transmit antennas to

the RUE k . $n_k(t)$ is the received noise at the RUE k with the distribution $\mathcal{CN}(0, \sigma^2)$, where σ^2 is the noise variance at RUEs. Eq. (2) suggests that both $\mathbf{v}_j(t)$ and $\mathbf{v}_{0,i}(t)$ should be carefully designed to suppress the intra-tier and inter-tier interference, respectively.

B. Queueing Model

Since the MBS in H-CRANs is mainly used to deliver the control signalling and provide seamless coverage with a low bit rate, for the ease of implementation, the beamformers of the MBS can be assumed to be fixed over a longer duration than the scheduling slot of RRHs, and thus the performance of MUEs can be assumed to remain stable if the inter-tier interference from RRHs is suppressed to a pre-defined threshold. Therefore, we can focus only on the performance optimization with queue stability for overall RRHs in H-CRANs under the condition that the queue stability of the MBS is guaranteed.

Suppose there are queues maintained for RUEs in H-CRANs which are represented by $\mathbf{Q}(t) = \{Q_k(t) | k = 1, \dots, K_R\}$, where $Q_k(t)$ denotes the queue backlog for RUE k at time slot t . The random multimedia traffic arrival for RUE k at the time slot t is denoted by $A_k(t)$, which is assumed to be independent and identically distributed (i.i.d.) over time slots with the peak arrival rate A_k^{max} . Define the set of $A_k(t)$ is as $\mathbf{A}(t) = \{A_k(t) | k = 1, \dots, K_R\}$, and the arrival rates of queues are $\boldsymbol{\lambda} = \mathbb{E}\{\mathbf{A}(t)\}$.

At each time slot, the arrival and departure rates of RUE k are $A_k(t)$ and $R_k(t)$, respectively. Therefore, $Q_k(t)$ evolves according to

$$Q_k(t+1) = \{Q_k(t) - R_k(t)\}^+ + A_k(t). \quad (3)$$

Considering the random and bursty characteristics of multimedia traffic arrivals and the QoS requirement of RUEs in H-CRANs, it is imperative to consider queue-aware resource allocation techniques. Therefore, to achieve this objective, the queue stability is defined as follows.

Definition 1: A discrete time process $Q(t)$ is mean-rate stable [28] if

$$\lim_{t \rightarrow \infty} \frac{\mathbb{E}\{|Q(t)|\}}{t} = 0. \quad (4)$$

Note that an absolute value of $Q(t)$ is used in the mean rate stability definition, which is useful for virtual queues those can be possibly negative.

C. Problem Formulation

To optimize the EE performance of H-CRANs, transmission rate and power consumption performance metrics should be jointly considered. According to (2), these two performance metrics are both presented as functions of the network-wide beamforming vector $\mathbf{v}_k(t)$ for RUE k :

- *Transmission Rate:* RUE k is scheduled at time slot t , i.e., $R_k(t)$ is nonzero if and only if its network-wide beamforming vector $\mathbf{v}_k(t)$ is nonzero. Based on the network-wide beamforming vector $\mathbf{v}_k(t)$ for RUE k and $\mathbf{v}_j(t)$ for RUE j , the signal-to-interference-plus-noise ratio (SINR) can be directly derived as $\mathbf{v}_k^H(t)\mathbf{h}_k(t) \left(\sum_{j=1, j \neq k}^{K_R} \mathbf{h}_k^H(t)\mathbf{v}_j(t)\mathbf{v}_j^H(t)\mathbf{h}_k(t) + \right.$

$\left. \phi_k(t) \right)^{-1} \mathbf{h}_k^H(t)\mathbf{v}_k(t)$. As a result, according to the Shannon capacity formula, the achievable transmission rate for RUE k at time slot t can be expressed as

$$R_k(t) = \log_2 \left(1 + \mathbf{v}_k^H(t)\mathbf{h}_k(t) \left(\sum_{j=1, j \neq k}^{K_R} \mathbf{h}_k^H(t)\mathbf{v}_j(t) \mathbf{v}_j^H(t)\mathbf{h}_k(t) + \phi_k(t) \right)^{-1} \mathbf{h}_k^H(t)\mathbf{v}_k(t) \right), \quad (5)$$

where $\phi_k(t) = \sum_{i=1}^{K_M} \mathbf{g}_{0,k}^H(t)\mathbf{v}_{0,i}(t)\mathbf{v}_{0,i}^H(t)\mathbf{g}_{0,k}(t) + \sigma^2$ can be assumed to remain constant in time slot t because the beamformers of the MBS are fixed.

- *Power Consumption:* Since the transmitted signals from RRHs to RUEs have unit variance, the radio frequency power consumption $P_n(t)$ of the n -th RRH in the time slot t depends only on the beamformer transmitting to the RUEs. Therefore, the power consumption for RRH n serving all potential K_R RUEs can be written as

$$P_n(t) = \sum_{k=1}^{K_R} \mathbf{v}_k^H(t)\mathbf{D}_n^H\mathbf{D}_n\mathbf{v}_k(t) + PC_n(t) + PF_n(t), \quad (6)$$

where $P_n(t)$ denotes the total power consumption of the n -th RRH, and $PC_n(t)$ and $PF_n(t)$ are the circuit power consumption and fronthaul power consumption of RRH n , respectively. Note that the circuit power of RRHs is negligible because the energy consumption of air conditioning is avoided. Since RRHs are connected to the BBU pool via optical fiber to alleviate fronthaul capacity constraints, the power consumption of the fronthaul is rather small compared with the transmit power of RRHs, and it can be neglected, too. Thus, the power consumption model can be reformulated as

$$P_n(t) = \sum_{k=1}^{K_R} \mathbf{v}_k^H(t)\mathbf{D}_n^H\mathbf{D}_n\mathbf{v}_k(t), \quad (7)$$

$$\bar{P}_n = \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}\{P_n(\tau)\},$$

where \bar{P}_n is the time average of $P_n(t)$.

The traditional EE metric is defined as the ratio of the weighted sum transmit rate to the corresponding weighted total energy consumption in units of bit/Hz/Joule, which is given by

$$\tilde{\eta}_{EE}(t) = \frac{\sum_{k=1}^{K_R} \omega'_k R_k(t)}{\sum_{n=1}^N \mu'_n P_n(t)}, \quad (8)$$

where $\omega'_k \geq 0$ and $\mu'_n \geq 0$ represent the transmit weight of user k and the power consumption weight of the n -th RRH, respectively.

Following [29] and [30], instead of directly maximizing $\tilde{\eta}_{EE}(t)$, we define an alternative form of EE, $\eta_{EE}(t)$, and aim at maximizing it.

Definition 2: To quantitatively capture the relative importance of transmit rate and power consumption, the weighted EE utility function $f(R_k(t), P_n(t))$ in terms of sum transmit rate and total energy consumption with varied weight factors is used [29] in this paper to denote the equivalent EE metric of overall RRHs in the time slot t as follows:

$$\begin{aligned} \eta_{EE}(t) &= f(R_k(t), P_n(t)) \\ &= \frac{\alpha}{K_R} \sum_{k=1}^{K_R} \omega_k R_k(t) - \frac{1-\alpha}{N} \sum_{n=1}^N \mu_n P_n(t), \end{aligned} \quad (9)$$

where $\alpha \in [0, 1]$ is a weighting factor representing the ratio of the transmit rate to the power consumption. Here, $\omega_k \geq 0$ (channels/bits) and $\mu_n \geq 0$ (W^{-1}) represent the transmit weight of user k and the power consumption weight of the n -th RRH, respectively.

Remark 1: (1) Note that $\eta_{EE}(t)$ can be used to characterize $\tilde{\eta}_{EE}(t)$ [29], cf. from Eq. (9) to Eq. (11) on Page 3]. (2) Since both $R_k(t)$ and $P_n(t)$ depend on the network-wide beamforming vector $\mathbf{v}_k(t)$, $\eta_{EE}(t)$ is mainly determined by $\mathbf{v}_k(t)$, and the optimization of $\eta_{EE}(t)$ is strictly related to $\mathbf{v}_k(t)$.

Note that the beamformers from RRHs to RUEs cause severe inter-tier interference to MUEs. Therefore, the inter-tier interference from RRHs to the k -th MUE should be constrained, which can be formulated as

$$\sum_{j=1}^{K_R} \mathbf{v}_j^H(t) \mathbf{g}_k(t) \mathbf{g}_k^H(t) \mathbf{v}_j(t) \leq \varphi_k, k \in \mathcal{K}_M, \forall t, \quad (10)$$

where $\mathbf{g}_k(t) \in \mathbb{C}^{N_{LR} \times 1}$ denotes the CSI matrix from all RRHs' transmit antennas to MUE k . Eq. (10) suggests that the overall inter-tier interference from adjacent RRHs should be suppressed to a pre-defined threshold by appropriately designing the network-wide beamforming vector $\mathbf{v}_j(t)$.

Meanwhile, the overall radio-over-fiber in-phase/quadrature (I/Q) fronthaul capacity for the n -th RRH is constrained by a capacity threshold F_n . Considering the compression techniques over the fronthaul, the fronthaul capacity is not exactly equal to the accumulated data rates but rather is a utility function of it [31]. A utility function should be used not only to present the linear relationship between the radio-over-fiber I/Q fronthaul capacity and the accumulated data rate, but also to incorporate the impact of the compression technique. Therefore, the fronthaul capacity constraint is expressed as

$$\begin{aligned} &g\left\{\sum_{k=1}^{K_R} \mathbb{1}\{\mathbf{v}_{n,k}^H(t) \mathbf{v}_{n,k}(t)\} R_k(t)\right\} \\ &= g\left\{\sum_{k=1}^{K_R} \mathbb{1}\{\mathbf{v}_k^H(t) \mathbf{D}_n^H \mathbf{D}_n \mathbf{v}_n(t)\} R_k(t)\right\} \leq F_n, n \in \mathcal{N}, \end{aligned} \quad (11)$$

where $g(\cdot)$ reflects the relationship between the accumulated data rate of radio access links and the radio-over-fiber I/Q fronthaul capacity under a given compression technique. Here, $\mathbb{1}\{x\}$ denotes the indicator function of set $R/\{0\}$ for $x \geq 0$:

$$\mathbb{1}\{x\} = \begin{cases} 0, & \text{if } x = 0 \\ 1, & \text{else} \end{cases}. \quad (12)$$

Let $C_n = g^{-1}(F_n)$; then the radio-over-fiber I/Q fronthaul capacity constraint can be expressed equivalently that the accumulated data rate is not beyond C_n , which can be written as

$$\begin{aligned} \sum_{k=1}^{K_R} \mathbb{1}\{\mathbf{v}_{n,k}^H(t) \mathbf{v}_{n,k}(t)\} R_k(t) &= \sum_{k=1}^{K_R} \mathbb{1}\{\mathbf{v}_k^H(t) \mathbf{D}_n^H \mathbf{D}_n \mathbf{v}_n(t)\} R_k(t) \\ &\leq C_n, n \in \mathcal{N}. \end{aligned} \quad (13)$$

When considering all constraints, including the average power consumption of each RRH expressed in (7), the queue stability expressed in (4), the inter-tier interference to MUEs expressed in (10), and the individual fronthaul capacity of each RRH expressed in (13), the maximization of the averaged weighted EE utility objective function for RRHs in H-CRANs can be formulated as the following stochastic optimization problem:

$$\begin{aligned} \max_{\{\mathbf{v}_k(t)\}} \quad &\overline{\eta_{EE}} = \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}\{\eta_{EE}(\tau)\} \\ \text{s.t.} \quad & \\ C1: \quad &\overline{P}_n \leq P_n^{avg}, n \in \mathcal{N}, \\ C2: \quad &\lim_{t \rightarrow \infty} \frac{\mathbb{E}\{|Q_k(t)|\}}{t} = 0, k \in \mathcal{K}_R, \\ C3: \quad &P_n(t) \leq P_n^{\max}, n \in \mathcal{N}, \\ C4: \quad &\sum_{j=1}^{K_R} \mathbf{v}_j^H(t) \mathbf{g}_k(t) \mathbf{g}_k^H(t) \mathbf{v}_j(t) \leq \varphi_k, k \in \mathcal{K}_M, \forall t, \\ C5: \quad &\sum_{k=1}^{K_R} \mathbb{1}\{\mathbf{v}_k^H(t) \mathbf{D}_n^H \mathbf{D}_n \mathbf{v}_n(t)\} R_k(t) \leq C_n, n \in \mathcal{N}. \end{aligned} \quad (14)$$

In (14), the constraint $C1$ ensures the long-term energy consumption of the n -th RRH under the predefined level where P_n^{avg} denotes the average power consumption threshold. $C2$ is the network stability constraint to guarantee a finite queue length for each queue. $C3$ is the energy-saving constraint for the n -th RRH where P_n^{\max} denotes the maximum transmit power of the n -th RRH. $C4$ is the constraint on interference from RRHs to MUEs, and $C5$ is the constraint on the fronthaul consumption for the n -th RRH.

Intuitively, the optimization objective function expressed in (14) with so many constraints is complex and cannot be directly solved. We also note that $C1$ and $C2$ in (14) are constraints on time averaged variables. Hence, they can be satisfied only if the BBU pool has the CSI and knowledge of queue backlogs at all time slots instantaneously, which is infeasible and unpractical. Fortunately, with Lyapunov optimization tool [28], the time-averaged constraints $C1$ and $C2$ can be transferred into instantaneous constraints, and the optimization function with the $C1$ and $C2$ constraints can be transformed into a queue mean-rate stable problem, which can be solved only based on the observed CSI and queue backlogs at each time slot.

III. DELAY-AWARE EE MAXIMIZATION

In this section, the optimization problem in (14) is reformulated as an equivalent sum-MSE minimization problem. Then a corresponding dynamic network-wide beamforming algorithm is proposed.

A. General Lyapunov Optimization

Before presenting the solution of the problem, we first give the following lemma to show how the average constraint can be transformed into a queue stability problem.

Lemma 1: Construct a virtual queue $H_n(t)$, the queue dynamics of which are

$$H_n(t+1) = \{H_n(t) - P_n^{avg} + P_n(t)\}^+. \quad (15)$$

Suppose $\mathbb{E}\{H_n(0)\} < \infty$, if the virtual queue $H_n(t)$ is mean-rate stable, the inequality $\bar{P}_n \leq P_n^{avg}$ can be satisfied.

Proof: Suppose that $H_n(t)$ is mean-rate stable; then we have $\lim_{T \rightarrow \infty} \frac{\mathbb{E}\{|H_n(T)|\}}{T} = 0$ with the probability 1 based on Definition 1. Summing (15) from $t = 0$ to $T-1$, we have that

$$\begin{aligned} \sum_{t=0}^{T-1} H_n(t+1) &= \sum_{t=0}^{T-1} \{H_n(t) - P_n^{avg} + P_n(t)\}^+ \\ &= \sum_{t=0}^{T-1} \{H_n(t) - P_n^{avg} \\ &\quad + \max\{P_n(t), P_n^{avg} - H_n(t)\}\} \\ H_n(T) - H_n(0) &= \sum_{t=0}^{T-1} \max\{P_n(t), P_n^{avg} - H_n(t)\} - TP_n^{avg} \\ &\geq \sum_{t=0}^{T-1} P_n(t) - TP_n^{avg}, \end{aligned} \quad (16)$$

holds for all time slots $t > 0$. Taking the sum-expectation operation and the limit as $T \rightarrow \infty$, we can conclude that

$$\lim_{T \rightarrow \infty} \frac{\mathbb{E}\{|H_n(T)|\}}{T} - \lim_{T \rightarrow \infty} \frac{\mathbb{E}\{|H_n(0)|\}}{T} = 0 \geq \bar{P}_n - P_n^{avg}. \quad (17)$$

Therefore, $\bar{P}_n \leq P_n^{avg}$ holds. ■

With Lemma 1, the constraint C1 in (11) is transformed into a queue stability problem by constructing a virtual queue $H_n(t)$ for each RRH n .

With the actual queues (3) and virtual queues (15), denote $\Theta(t) = [\mathbf{Q}(t), \mathbf{H}(t)]$ as the combined matrix of all the actual and virtual queues, where $\mathbf{H}(t) = \{H_n(t) | n = 1, \dots, N\}$, the Lyapunov function is defined as a scalar metric of queue congestion:

$$L(\Theta(t)) \triangleq \frac{1}{2} \left\{ \sum_{k=1}^{K_R} Q_k^2(t) + \sum_{n=1}^N H_n^2(t) \right\}. \quad (18)$$

The Lyapunov drift is introduced to push the Lyapunov function to a lower congestion state and keep queues stable, which is defined as

$$\Delta(\Theta(t)) \triangleq \mathbb{E}\{L(\Theta(t+1)) - L(\Theta(t))\}. \quad (19)$$

In terms of Lyapunov optimization, the underlying objective of optimal network-wide beamformer design is to minimize an

infimum bound on the drift-plus-penalty expression in each time slot:

$$\Delta(\Theta(t)) - V\mathbb{E}\{\eta_{EE}(t)|\Theta(t)\}, \quad (20)$$

where V is a non-negative parameter controlling the tradeoff between the average weighted EE performance and average queue backlog. For simplicity, this parameter is termed as EE-delay tradeoff in the following discussions.

With the queue dynamics of $Q_k(t)$ and $H_n(t)$ presented in (3) and (15), respectively, and the definition of Lyapunov drift in (19), the following lemma holds.

Lemma 2: At any time slot t , with the observed queue state and CSI, and under any network-wide beamformer control decision, the drift-plus-penalty satisfies the following inequality:

$$\begin{aligned} &\Delta(\Theta(t)) - V\mathbb{E}\{\eta_{EE}(t)|\Theta(t)\} \\ &\leq B - V\mathbb{E}\{\eta_{EE}(t)|\Theta(t)\} \\ &\quad + \sum_{n=1}^N H_n(t)\mathbb{E}\{P_n(t) - P_n^{avg}|\Theta(t)\} \\ &\quad + \sum_{k=1}^{K_R} Q_k(t)\mathbb{E}\{A_k(t) - R_k(t)|\Theta(t)\}, \end{aligned} \quad (21)$$

where $B > 0$ is a finite constant which satisfies (22) for $\forall t$:

$$\begin{aligned} B \geq &\frac{1}{2}\mathbb{E}\left\{\sum_{k=1}^{K_R} (A_k^2(t) + R_k^2(t))|\Theta(t)\right\} \\ &+ \frac{1}{2}\mathbb{E}\left\{\sum_{n=1}^N (P_n(t) - P_n^{avg}(t))^2|\Theta(t)\right\}. \end{aligned} \quad (22)$$

Proof: See Appendix A. ■

To push the underlying objective (20) to its minimum, a proper network-wide beamformer $\mathbf{v}_k(t)$ is chosen to greedily minimize the drift-plus-penalty in (20). As a result, a strategy is proposed herein to minimize the right-hand-side (R.H.S) of the inequality of drift-plus-penalty in (21) based on the observed QSI and CSI at each time slot t instead of minimizing (20) directly. Based on the concept of opportunistically minimizing an expectation, this is accomplished by greedily minimizing as follows:

$$\min_{\{\mathbf{v}_k(t)\}} \left[\sum_{n=1}^N H_n(t)P_n(t) - \sum_{k=1}^{K_R} Q_k(t)R_k(t) - V\eta_{EE}(t) \right]. \quad (23)$$

For notational simplicity, $X_n(t)$ and $Y_k(t)$ are denoted by

$$\begin{aligned} X_n(t) &= H_n(t) + \frac{V(1-\alpha)\mu_n}{N}, \\ Y_k(t) &= Q_k(t) + \frac{V\alpha\omega_k}{K_R}, \end{aligned} \quad (24)$$

respectively. With the constraints C1 and C2 incorporated into the objective function (23), the optimization problem (14) can be rewritten as

$$\begin{aligned} &\min_{\{\mathbf{v}_k(t)\}} \sum_{n=1}^N X_n(t)P_n(t) - \sum_{k=1}^{K_R} Y_k(t)R_k(t) \\ &s.t. \quad C3, C4, C5, \end{aligned} \quad (25)$$

where $X_n(t)$ and $Y_k(t)$ can be calculated by the observed QSI at time slot t . $P_n(t)$ and $R_k(t)$ are based on the CSI at time slot t and the beamforming vector $\mathbf{v}_k(t)$. Thus, a network-wide cooperative beamformer design algorithm is proposed in the next subsection to solve problem (25).

B. Beamformer Design Algorithm

The optimization problem (25) is non-convex, which is difficult to solve directly. To present a solution of (25), the performance of control actions to obtain a local minimum which is within an additive constant of the infimum is analyzed. Therefore, in the following content, the definition of *C-additive approximation* [28] is first introduced, based on which the locally optimal solution of problem (25) is analyzed. Finally, a network-wide beamformer design algorithm is proposed.

Definition 3: For a given constant $C \geq 0$, a *C-additive approximation* of the drift-plus-penalty algorithm is to choose an action that yields a conditional expected value on the right-hand-side of the drift-plus-penalty (given $\Theta(t)$) at time slot t , that is within a constant C from the infimum over all possible control actions.

Based on *Definition 3*, a local optimal beamformer algorithm can be designed. Note that in the constraint $C5$, the indicator function can be equivalently expressed as a scalar ℓ_0 -norm, which is the number of nonzero entries in a vector. The indicator function can be approximated by a convex re-weighted ℓ_1 -norm [32] [33], i.e.,

$$\mathbb{1}\{\mathbf{v}_k^H(t)\mathbf{D}_n^H\mathbf{D}_n\mathbf{v}_k(t)\} = \beta_n^k(t)\mathbf{v}_k^H(t)\mathbf{D}_n^H\mathbf{D}_n\mathbf{v}_k(t), \quad (26)$$

where $\beta_n^k(t)$ is updated iteratively according to $\beta_n^k(t) = \frac{1}{\mathbf{v}_k^H(t)\mathbf{D}_n^H\mathbf{D}_n\mathbf{v}_k(t) + \kappa}$ with a small constant regularization factor $\kappa > 0$ and \mathbf{v}_k from the previous iteration. Since it is still difficult to solve a problem that involves $R_k(t)$ in both the objective function and the constraints, an iterative scheme is used in which the fixed value of $R_k(t)$ from the previous iteration is adopted here. Thus, the optimization problem can be written as

$$\begin{aligned} \min_{\{\mathbf{v}_k(t)\}} \quad & \sum_{n=1}^N X_n(t)P_n(t) - \sum_{k=1}^{K_R} Y_k(t)R_k(t) \\ \text{s.t.} \quad & C3, C4, \\ & C6 : \sum_{k=1}^{K_R} \beta_n^k \mathbf{v}_k^H(t)\mathbf{D}_n^H\mathbf{D}_n\mathbf{v}_k(t)\tilde{R}_k(t) \leq C_n, \end{aligned} \quad (27)$$

where $\tilde{R}_k(t)$ in $C6$ is the rate of the previous iteration. Obviously, the approximated problem (27) is still non-convex, while it can be reformulated as an equivalent WMMSE problem to achieve a local optimum via the *C-additive approximation* of the drift-plus-penalty algorithm. Inspired by the equivalence between the weighted sum rate (WSR) maximization and WMMSE [34]–[36] for the MIMO channel, the generalized WMMSE equivalence established in [35] is extended to solve the problem (27) in the H-CRAN scenario. We state this equivalence as follows.

Proposition 1: The problem (27) has the same optimal solution as the following WMMSE problem:

$$\begin{aligned} \min_{\{w_k(t), u_k(t), \mathbf{v}_k(t)\}} \quad & \sum_{k=1}^{K_R} Y_k(t) \{w_k(t)e_k(t) - \log w_k(t)\} \\ & + \sum_{n=1}^N X_n(t) \sum_{k=1}^{K_R} \mathbf{v}_k^H(t)\mathbf{D}_n^H\mathbf{D}_n\mathbf{v}_k(t), \\ \text{s.t.} \quad & C3, C4, C6. \end{aligned} \quad (28)$$

where $w_k(t)$ denotes the mean-square error (MSE) weight for user k at time slot t , and $e_k(t)$ is the corresponding MSE defined as

$$\begin{aligned} e_k(t) & \triangleq \mathbb{E} \left\{ (u_k(t)y_k(t) - s_k(t))^2 \right\} \\ & = u_k^H(t) \left(\sum_{j=1}^{K_R} \mathbf{v}_j^H(t)\mathbf{h}_k(t)\mathbf{h}_k^H(t)\mathbf{v}_j(t) \right) u_k(t) \\ & \quad + u_k^H(t) \left(\sum_{i=1}^{K_M} \mathbf{v}_{0,i}^H(t)\mathbf{g}_{0,k}(t)\mathbf{g}_{0,k}^H(t)\mathbf{v}_{0,i}(t) \right) u_k(t) \\ & \quad - 2\mathbf{Re}\{u_k(t)\mathbf{h}_k^H(t)\mathbf{v}_k(t)\} + \sigma^2\mathbf{Re}\{u_k^H(t)u_k(t)\} + 1, \end{aligned} \quad (29)$$

under the receiver $u_k(t) \in \mathbb{C}$.

Based on the equivalent WMMSE problem (28) which is convex with respect to each of the individual optimization variables, the averaged weighted EE utility objective maximization problem (14) can be solved. This crucial observation allows the problem (14) to be solved efficiently through the block coordinate descent method by iterating among $\mathbf{v}_k(t)$, $u_k(t)$, and $w_k(t)$:

- The optimal MSE weight $w_k(t)$ under the fixed $\mathbf{v}_k(t)$ and $u_k(t)$ is given by

$$w_k^{opt}(t) = e^{-1}(t). \quad (30)$$

- The optimal receiver $u_k(t)$ under the fixed $\mathbf{v}_k(t)$ and $w_k(t)$ is given by

$$u_k^{opt}(t) = \mathbf{h}_k^H(t)\mathbf{v}_k(t) \left\{ \sum_{j=1}^{K_R} \mathbf{v}_j^H(t)\mathbf{h}_k(t)\mathbf{h}_k^H(t)\mathbf{v}_j(t) + \phi(t) \right\}^{-1}. \quad (31)$$

- The optimization problem for finding the optimal transmit network-wide beamformer $\mathbf{v}_k(t)$ under the fixed $w_k(t)$ and $u_k(t)$ is

$$\begin{aligned} \min_{\{\mathbf{v}_k(t)\}} \quad & \sum_{k=1}^{K_R} \mathbf{v}_k^H(t) \left(\sum_{j=1}^{K_R} Y_j(t)w_j(t)u_j^H(t)\mathbf{h}_j(t) \right. \\ & \left. \mathbf{h}_j^H(t)u_j(t) + \sum_{n=1}^N X_n(t)\mathbf{D}_n^H\mathbf{D}_n \right) \mathbf{v}_k(t) \\ & - 2 \sum_{k=1}^{K_R} Y_k(t)w_k(t)\mathbf{Re}\{u_k(t)\mathbf{h}_k^H(t)\mathbf{v}_k(t)\} \\ \text{s.t.} \quad & C3, C4, C6. \end{aligned} \quad (32)$$

The optimization objective function in (32) is a quadratically constrained quadratic programming (QCQP) problem and can be solved using a standard convex optimization solver such as the Matlab software for disciplined convex programming (CVX) [37].

The above proposed WMMSE approach for solving the original optimization problem (14) can be summarized in **Algorithm 1**.

Algorithm 1 Averaged weighted EE maximization with per-RRH power and interference constraints at time slot t .

Require: Initial network-wide beamforming vector $\mathbf{v}_k(t)$ and corresponding $\eta_{EE}(t)$, and the precision κ ;

Ensure: Calculate the optimal $\mathbf{v}_k^*(t)$ and corresponding $\eta_{EE}^*(t)$.

1: **repeat**

2: **Update** $\mathbf{v}_k^*(t) = \mathbf{v}_k(t)$ and $\eta_{EE}^*(t) = \eta_{EE}(t)$;

3: With $\mathbf{v}_k(t)$ fixed, compute the MMSE receiver $u_k(t)$ and the corresponding MSE $e_k(t)$ according to (31) and (29);

4: Update the MSE weight $w_k(t)$ according to (30);

5: Find the optimal transmit network-wide beamformer $\mathbf{v}_k(t)$ under fixed $u_k(t)$ and $w_k(t)$, by solving the QCQP problem (32);

6: Compute the achievable rate $R_k(t)$ and energy consumption $P_n(t)$ according to (5) and (7), respectively;

7: Compute the EE function $\eta_{EE}(t)$;

8: **Update** $\beta_n^k(t)$ and $\bar{R}_k(t)$.

9: **until** $|\eta_{EE}^*(t) - \eta_{EE}(t)| \leq \kappa |\eta_{EE}^*(t)|$.

Note that **Algorithm 1** cannot be guaranteed to converge to the global optimum, while it can quickly converge to a local optimum. The random initial point can approach a local optimum through **Algorithm 1** with a substantial number of iterations. To decrease the number of iterations and approach a local optimal solution quickly, it is critical to choose proper initialization points with reasonable approaches such as the interference alignment initialization proposed in [38].

Remark 2: **Algorithm 1** is based on the block coordinate descent method. In this case, the computational complexity of Step 2 in **Algorithm 1** is $O(K_R^2 N L_R)$, mainly due to the receive covariance matrix computation in (29) and (31). With the MSE $e_k(t)$ obtained from Step 2, the additional computational complexity for Step 3 to update all MSE weights $w_k(t)$ is only $O(K)$. Step 4 requires solution of the QCQP problem, which is the largest part of the computational complexity in **Algorithm 1**. The total number of variables in the QCQP problem is $K_R L_R N$ and the computation complexity of using the CVX method to solve such a QCQP problem is approximately $O((K_R L_R N)^{3.5})$. Step 5 has the same computational complexity as computing the MSE, and the computational complexity of the remaining steps is $O(K)$.

It is noted that the complexity of solving such a QCQP problem is related to the number of potential transmit antennas $L_R N$ serving each user and the total number of users K_R to be considered in each iteration. Thus, to improve the efficiency of

Algorithm 1 in each iteration, it is practical to reduce the number of potential transmit antennas $L_R N$ and the total number of users K_R . This can be done by iteratively removing the n -th RRH from the k -th RUE's candidate cluster once the transmit power from the n -th RRH to the k -th RUE, i.e., $\mathbf{v}_{n,k}^H \mathbf{v}_{n,k}$, is below a certain threshold, or checking the achievable RUE rate R_k and energy consumption P_n iteratively and ignoring those RUEs with negligible rates during the next iteration. Such solutions can reduce the number of needed iterations and decrease the complexity of **Algorithm 1**.

IV. PERFORMANCE ANALYSIS

In this section, the average performance of the weighted EE utility function and the queue length bound achieved by the proposed **Algorithm 1** are introduced, which leads to an EE-delay tradeoff.

Before presenting the theoretical performance of the proposed **Algorithm 1**, we introduce the following assumption under the given channel condition and network-wide beamformer design algorithm:

$$R_k(t) \leq R_k^{max}, \mathbb{E}\{R_k^2(t)\} \leq (R_k^{max})^2, \quad (33)$$

$$\mathbb{E}\{P_n(t) - P_n^{av}\} \leq \gamma, \quad (34)$$

$$\mathbb{E}\{\eta_{EE}(t)\} \leq \eta_{EE}^{max}, \quad (35)$$

where R_k^{max} , γ and η_{EE}^{max} are finite constants. The assumptions (33) and (34) are reasonable in realistic systems since the data rate of each RUE and the power allocation of each RRH are constrained. The assumption (35) strictly depends on the boundedness assumptions of (33) and the power constraints.

Definition 4: The network capacity region is the set Λ of non-negative rate vectors $\boldsymbol{\lambda}$, satisfying:

$$\epsilon_{max}(\boldsymbol{\lambda}) \geq 0, \quad (36)$$

where $\epsilon_{max}(\boldsymbol{\lambda})$ is the maximum value of ϵ which satisfies $\lambda_k + \epsilon \leq \bar{R}_k$, $k \in \mathcal{K}_R$.

Lemma 3: Suppose that $\boldsymbol{\lambda}$ is strictly interior to the capacity region Λ , and let ϵ be a positive value such that $\epsilon < \epsilon_{max}(\boldsymbol{\lambda})$. If constraints are feasible, then for any $\theta > 0$, there exists an algorithm that makes independent, stationary and randomized decisions about the network-wide beamformer at each time slot based only on the observed network state, which satisfies

$$\begin{aligned} \mathbb{E}\{A_k(t) - R_k^*(t) | \Theta(t)\} &= \mathbb{E}\{A_k(t) - R_k^*(t)\} \leq -\epsilon, \\ \mathbb{E}\{P_n(t) - P_n^{av} | \Theta(t)\} &\leq \theta, \\ \mathbb{E}\{\eta_{EE}^*(t) | \Theta(t)\} &\geq \eta_{EE}^{opt} - \theta, \end{aligned} \quad (37)$$

where $\eta_{EE}^*(t)$, $y_n^*(t)$ and $R_m^*(t)$ are corresponding results under the stationary algorithm, and η_{EE}^{opt} is the theoretical optimal value of $\bar{\eta}_{EE}$ under constraints C1, C2, C3, C4 and C5 in (14).

The detailed proof for *Lemma 3* are omitted for simplicity as a similar proof can be found in [28].

A. Stability of Queues

In Section III, we proposed a beamformer design algorithm utilizing Lyapunov optimization technique, and the constraints

$C1$, $C2$ are incorporated into the process of Lyapunov optimization. *Theorem 1* shows that constraints $C1$, $C2$ are guaranteed under the proposed **Algorithm 1**.

Theorem 1: Suppose that $\mathbb{E}\{L(\Theta(0))\} < \infty$ and the multimedia traffic arrival rate λ is in within the network capacity region, then under the proposed beamformer design algorithm, all the actual queues and virtual queues are mean-rate stable.

Proof: See Appendix B. ■

Theorem 1 shows that constraints $C1$ and $C2$ are satisfied under **Algorithm 1** according to *Definition 1* and *Lemma 1*, respectively.

B. Average weighted EE Performance

The average weighted EE performance obtained by **Algorithm 1** is given by *Theorem 2*.

Theorem 2: Utilizing the proposed optimization solution, for any $V > 0$, the gap between the average weighted EE and the optimum is within $\mathcal{O}(1/V)$:

$$\overline{\eta_{EE}} \geq \eta_{EE}^{opt} - \frac{B+C}{V}, \quad (38)$$

where C is a constant gap between the local optimum obtained by the proposed dynamic network-wide beamformer design algorithm and the infimum of the R.H.S of (21).

Proof: See Appendix C. ■

As the optimal solution of the optimization problem (14) with different non-trivial constraints is difficult to obtain in practice. *Theorem 2* suggests that a near-to-optimal solution can be obtained arbitrarily close to the optimum by adjusting the control parameter V . That is, if V is sufficiently large, the average weighted EE performance can be pushed arbitrarily close to the optimum, which is more realistic than attempting to achieve the optimal with high complexity.

C. Queueing Bounds

To evaluate the constant queueing bound of the average queue length for the proposed dynamic network-wide beamformer design algorithm, the following theorem can be used.

Theorem 3: Assume that the network-wide beamformer of each RRH and the queue dynamics are determined by the proposed dynamic **Algorithm 1**. Then, the average queue bound satisfies

$$\overline{Q} = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{k=1}^{K_R} \mathbb{E}\{Q_k(t)\} \leq \frac{B+C+V(\eta_{EE}^{max} - \eta_{EE}^{opt})}{\epsilon}. \quad (39)$$

Proof: See Appendix D. ■

Theorem 3 shows that the average queue length is bounded by a deterministic upper backlog bound which increases linearly with V .

Remark 2: *Theorem 2* combined with *Theorem 3* show that the proposed dynamic network-wide beamformer design algorithm achieves an $[\mathcal{O}(1/V), \mathcal{O}(V)]$ tradeoff between the average weighted EE performance and queue backlogs, which leads to an EE-delay tradeoff for a given arrival rate according to Little's Theorem [39]. With an increase of the control parameter V , the achieved weighted EE performance becomes better at the cost of incurring the larger queueing delay.

Therefore, it is important to choose a proper V to obtain the required performance and QoS in realistic H-CRANs.

V. SIMULATION RESULTS AND ANALYSIS

To evaluate the performance of the proposed optimal network-wide cooperative beamformer design algorithm in H-CRANs, we consider one radio resource block for all RRHs and MBSs. When more radio resource blocks are used for each RRH and MBS, the radio resource block allocation algorithms that adapt to the time-varied radio channel and dynamic traffic arrival described in [10]–[12] can be directly used along with the proposal in this paper. To decrease the high complexity and reduce the simulation time, a small-scale H-CRAN system consisting of one MBS and 2 RRHs is considered with the assumed simulation parameters shown in Table I. Since only one radio resource block is considered, only one RUE can be served in each RRH. As a result, it is assumed that $K_R = 4$ and $K_M = 4$. Note that similar simulation results to those presented below can be achieved for a large-scale H-CRAN system consisting of more RRHs and MBSs.

TABLE I
SIMULATION PARAMETERS

Num. of (MBSs, RRHs, MUEs, RUEs)	(1, 2, 4, 4)
Num. of antennas / (MBS, RRH)	(2, 2)
Noise power spectral density	-174 dBm/Hz
Path loss exponent for transmission from BS to UE	4
Small-scale fading	Rayleigh fading
Maximum transmit power of RRH	0.22W
Average transmit power constraint of RRH	0.2 W
Transmit power of MBS	20W

A. Queue Stability Evaluation

To evaluate the queue stability achieved by the proposed dynamic network-wide beamformer design algorithm, we take the user queues at the arrival rate $\lambda = 4.2$ bit/slot/Hz as sample queues. The user average queue length against t under several V is shown in Fig. 2. It can be observed that the average queue length first increases with t and then fluctuates around certain fixed values. Larger V leads to larger stable values which directly follows with *Theorem 3*.

B. The EE-Delay Tradeoff

The average queue backlog achieved by the proposed dynamic network-wide beamformer design algorithm shown in Fig. 3 grows linearly in $\mathcal{O}(V)$ under given multimedia traffic arrival rate λ , which consolidates *Theorem 3*. Under the same V , the queue length differs when the traffic arrival rate λ changes, since different arrival rates cause different amounts of power consumption. Intuitively, a larger λ leads to higher average power consumption since more power is required to transmit data arrivals and avoid queuing congestion, which is supported by Fig. 4.

Fig. 4 is shown to evaluate the average power consumption as a function of V . It is observed that the larger the multimedia

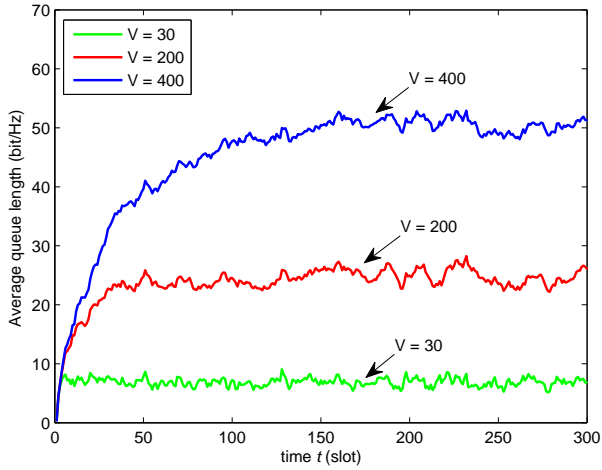


Fig. 2. Queue length versus simulation time length

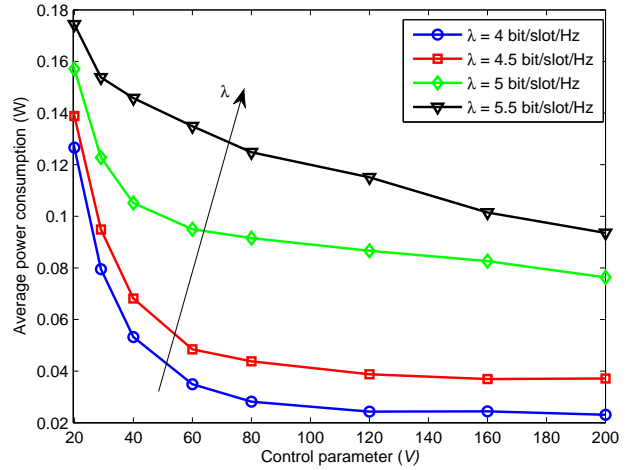


Fig. 4. Average power consumption versus V

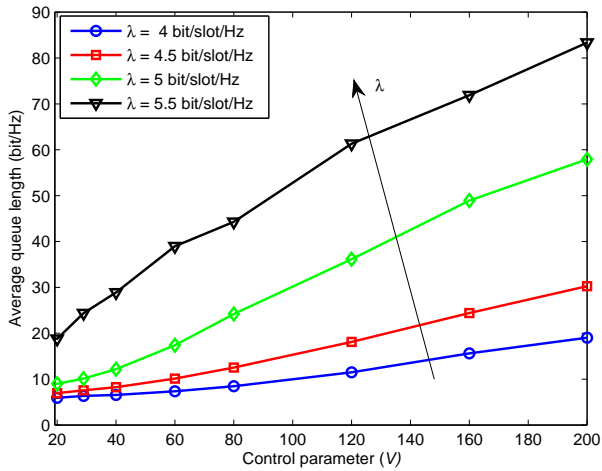


Fig. 3. Average queue length versus V

the logarithmic rate-power function has the characteristic of diminishing slope .

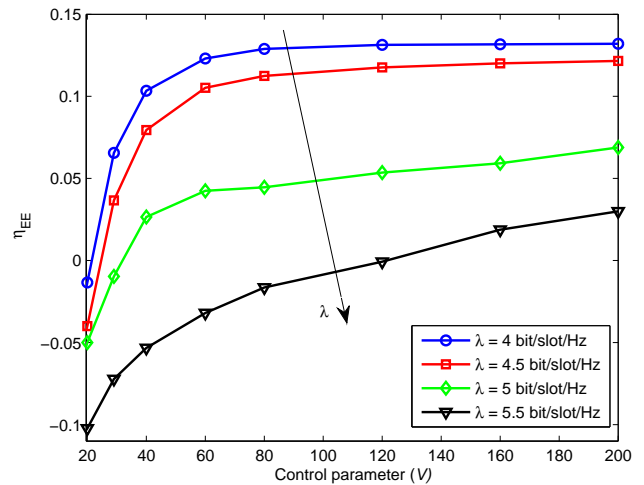


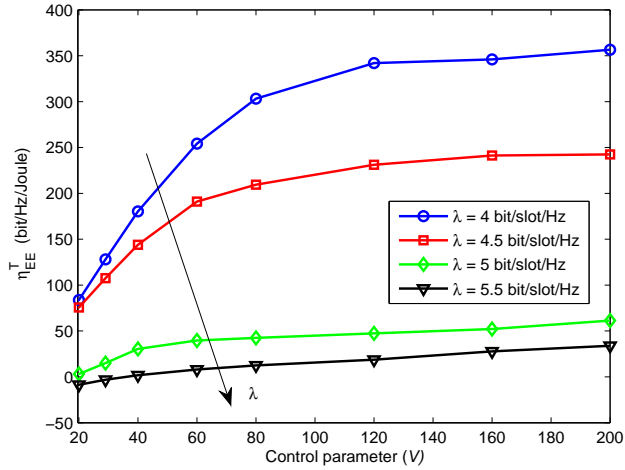
Fig. 5. Average weighted EE performance versus V

traffic arrival rate λ , the bigger the average power consumption. This is due to the fact that it is required for the system to consume more power to timely transmit more multimedia traffic arrivals. Meanwhile, the average power consumption decreases as V increases for a given λ . This can be explained by the fact that a larger V implies that the system emphasizes the average weighted EE more, but an increase in transmit power does not result in a proportional increase in transmit rate due to the diminishing slope of the logarithmic rate-power function. Therefore, it is necessary to decrease the transmit power to improve the average weighted EE performance.

In Fig. 5, the average weighted EE performance versus the parameter V for different user arrival rates λ is evaluated to support *Theorem 2*. The average weighted EE performance increases with V for any given arrival rate λ , which can be intuitively understood by the fact that greater emphasis is placed on the average weighted EE more when V increases. The lower the multimedia traffic arrival rate λ is, the higher the average weighted EE performance under a given control parameter V will be. This happens because both transmit rate and power consumption decrease with decreasing λ and the

To emphasize the efficiency and usefulness of the proposed average weighted EE performance metric η_{EE} , the traditional EE performance metric $\tilde{\eta}_{EE}$ defined in (8) is illustrated in Fig. 6 as a baseline. In Fig. 6, the average $\tilde{\eta}_{EE}$ grows at the rate of $O(1/V)$. The performance of $\tilde{\eta}_{EE}$ becomes better with a larger arrival data rate λ . It can be observed that the tendency of the average performance of $\tilde{\eta}_{EE}$ is similar to the average performance of the proposed EE metric η_{EE} , which indicates that the proposed EE metric η_{EE} is valid for use as the EE measurement in H-CRANs.

To make the tradeoff between the average weighted EE and the average queuing delay clear, the relationship between the average weighted EE and queuing delay is illustrated in Fig. 7 versus the parameter V . It can be observed that a larger V leads to a better average weighted EE performance but at the cost of incurring a larger queuing delay and vice versa. Thus, the proposed network-wide cooperative beamformer design algorithm provides an advanced method to flexibly balance


 Fig. 6. The traditional EE performance metric versus V

the average weighted EE performance and the queuing delay.

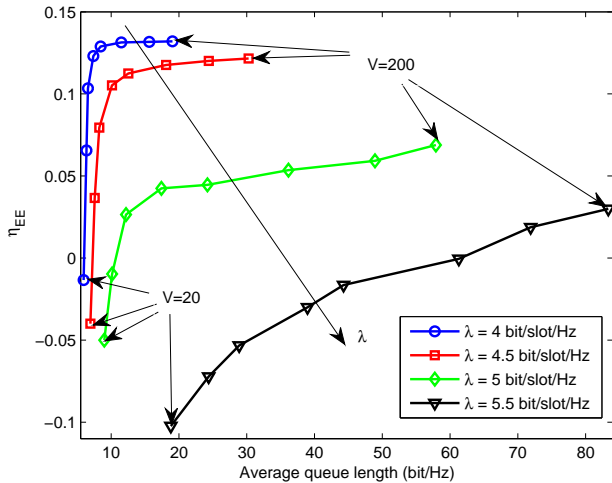
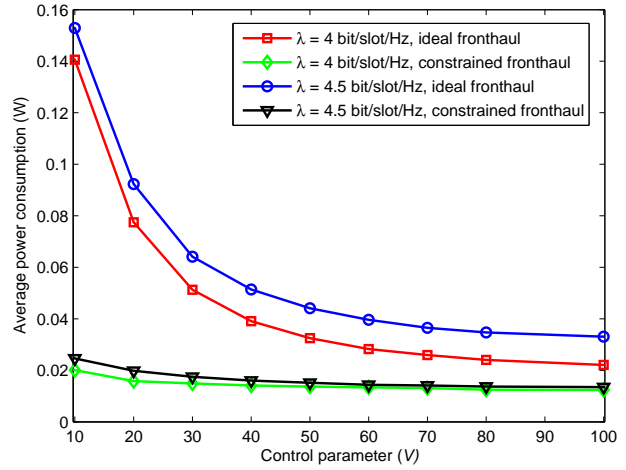
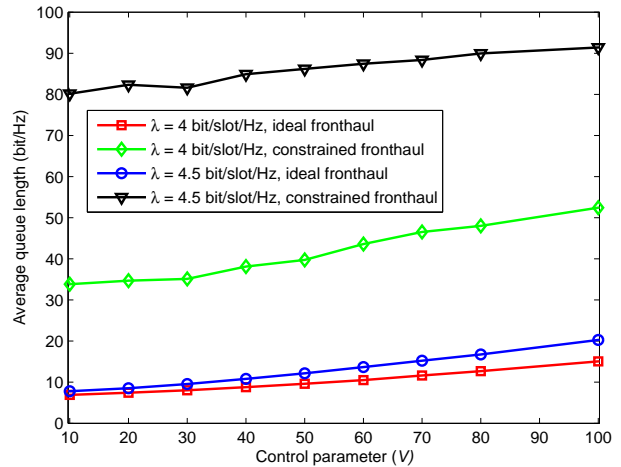


Fig. 7. Average weighted EE performance versus average queue length

C. Impact of fronthaul constraint on average weighted EE

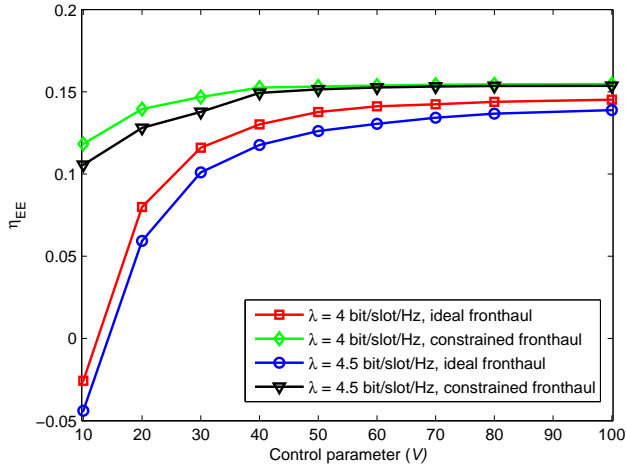
To evaluate the impact of the constrained fronthaul on the average weighted EE performance in H-CRANs, we compare the proposed network-wide cooperative beamformer design algorithm under constrained fronthaul $C_n = 6$ bit/Hz with the solution that maximizes EE with ideal fronthauls. As shown in Fig. 8, compared with the fronthaul constraint under the proposed dynamic network-wide beamformer design algorithm under the same arrival rate, the average power consumption becomes larger if the fronthaul constraint is not considered. This happens because the rising average power consumption leads to an increase of $\|\mathbf{v}_k\|$ in (1).

Meanwhile, under the same multimedia traffic arrival rate, the average queue length with the fronthaul constraint is larger than without the fronthaul constraint, which is illustrated in Fig. 9. The rational explanation is that the fronthaul constraint limits the transmission rate and causes more congestion, which results in larger average queue length.

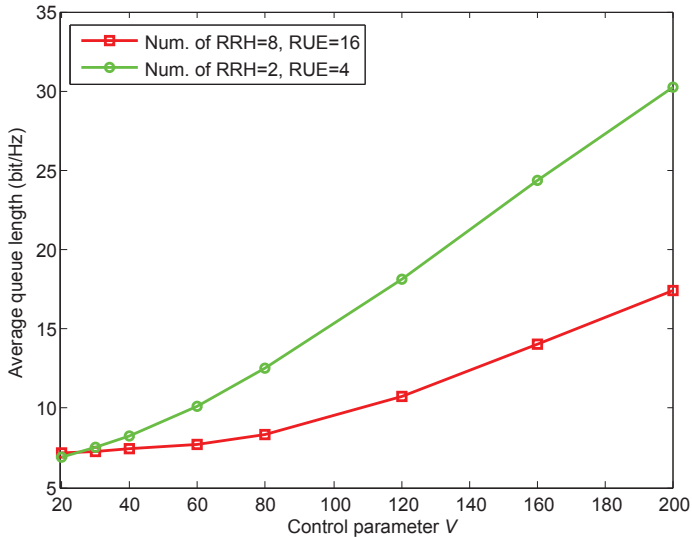

 Fig. 8. Average power consumption vs. V

 Fig. 9. Average queue length versus V

The average weighted EE performance under the ideal and constrained fronthaul are compared in Fig. 10, where the average performance of η_{EE} under the fronthaul constraint is better than that under the ideal fronthaul situation with the same multimedia traffic rate. Combined with Fig. 9, it can be concluded that fronthaul constraint leads to less power consumption and better energy efficiency performance but at the cost of incurring larger queuing delay.

Remark 3: In a practical H-CRAN, each RUE is associated with only a small number of adjacent RRHs. Thus, to decrease the computational and operational complexity of the simulation, we first conduct the simulation in a small area with 2 RRHs and 4 RUEs. In fact, when the simulated network size increases, similar simulation results are achieved though the simulation time becomes long. To illustrate this, we conduct another simulation configuration with 8 RRHs and 16 RUEs in the same concerned area. The obtained average queue lengths for these two simulation configurations are compared in Fig. 11. It can be observed that the average queue lengths under these two simulation configurations are nearly equal, both grow linearly at $O(V)$. The average performance of η_{EE} is

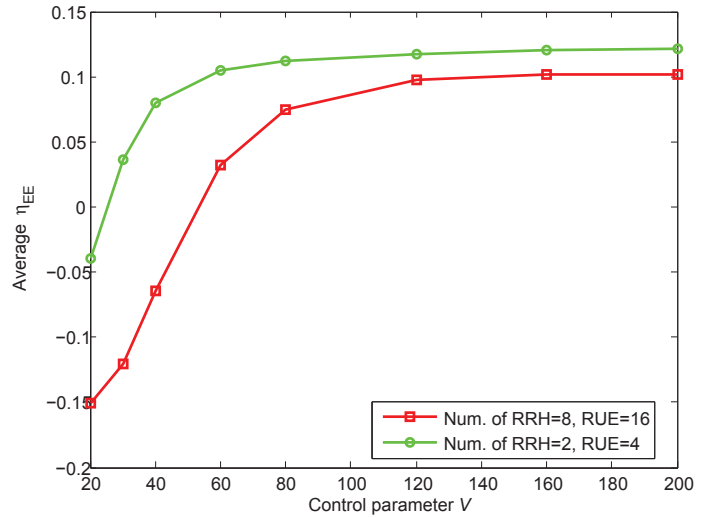

 Fig. 10. Average weighted EE performance versus V

compared in Fig. 12, which grows at the same tendency with V in Fig. 11. Therefore, we can conclude that the growth rate of the considered performance with the same simulation area are almost stable when the simulation network size becomes large.


 Fig. 11. Average power consumption versus V

VI. CONCLUSION

In this paper, to make the average energy efficiency arbitrarily close to the optimum and make each user's queue stable in multimedia H-CRANs, an average weighted EE performance metric has been proposed. Based on the advanced EE performance metric, a dynamic network-wide beamformer design algorithm based on the Lyapunov optimization framework has been proposed, which takes the average and instantaneous power constraints and the interference constraints into account. This network-wide beamformer design algorithm can be used to solve the non-convex average weighted EE performance optimization problem via a general weighted minimum mean square error (WMMSE) approach. An $[\mathcal{O}(1/V), \mathcal{O}(V)]$ EE-delay tradeoff is finally achieved by the proposed algorithm,


 Fig. 12. Average power consumption versus V

which is verified by both the mathematical analysis and numerical evaluations. The results have shown that the optimal average weighted EE performance under varied queue lengths strictly depends on the control parameter V . Furthermore, the fronthaul constraint has a significant impact on the average weighted EE performance. In realistic multimedia H-CRANs, the optimal V should be pre-selected to optimize the average weighted EE performance with both ideal and constrained fronthaul under the given multimedia queuing delay configuration.

APPENDIX A PROOF OF LEMMA 2

By squaring both sides of (3), the following inequality can be obtained

$$\begin{aligned}
 Q_k^2(t+1) &\leq Q_k^2(t) + R_k^2(t) + A_k^2(t) - 2Q_k(t)R_k(t) \\
 &\quad + 2A_k(t)\{Q_k(t) - R_k(t)\}^+ \\
 &\leq Q_k^2(t) + R_k^2(t) + A_k^2(t) + 2Q_k(t)\{A_k(t) - R_k(t)\}.
 \end{aligned} \tag{40}$$

Summing (40) over $k \in \{1, 2, \dots, K_R\}$, we obtain

$$\begin{aligned}
 &\frac{1}{2} \left\{ \sum_{k=1}^{K_R} Q_k^2(t+1) - \sum_{k=1}^{K_R} Q_k^2(t) \right\} \\
 &\leq \frac{1}{2} \sum_{k=1}^{K_R} \{R_k^2(t) + A_k^2(t)\} - \sum_{k=1}^{K_R} Q_k(t)\{R_k(t) - A_k(t)\}.
 \end{aligned} \tag{41}$$

Similarly, for virtual queues $H_n(t)$, we have

$$\begin{aligned}
 &\frac{1}{2} \left\{ \sum_{n=1}^N H_n^2(t+1) - \sum_{n=1}^N H_n^2(t) \right\} \\
 &\leq \frac{1}{2} \sum_{n=1}^N \{P_n(t) - P_n^{avg}(t)\}^2 + \sum_{n=1}^N H_n(t)\{P_n(t) - P_n^{avg}(t)\}.
 \end{aligned} \tag{42}$$

Summing (41) and (42) and taking expectations of both

sides to yield

$$\begin{aligned}
 & \mathbb{E}\{L(\Theta(t+1)) - L(\Theta(t)|\Theta(t))\} \\
 & \leq \frac{1}{2} \sum_{k=1}^{K_R} \mathbb{E}\{R_k^2(t) + A_k^2(t)|\Theta(t)\} \\
 & \quad + \sum_{k=1}^{K_R} Q_k(t) \mathbb{E}\{A_k(t) - R_k(t)|\Theta(t)\} \\
 & \quad + \frac{1}{2} \sum_{n=1}^N \mathbb{E}\{P_n(t) - P_n^{avg}(t)|\Theta(t)\}^2 \\
 & \quad + \sum_{n=1}^N H_n(t) \mathbb{E}\{P_n(t) - P_n^{avg}(t)|\Theta(t)\}.
 \end{aligned} \tag{43}$$

Subtracting $V\mathbb{E}\{\eta_{EE}(t)|\Theta(t)\}$, we have

$$\begin{aligned}
 & \Delta(\Theta(t)) - V\mathbb{E}\{\eta_{EE}(t)|\Theta(t)\} \\
 & \leq B - V\mathbb{E}\{\eta_{EE}(t)|\Theta(t)\} \\
 & \quad + \sum_{k=1}^{K_R} Q_k(t) \mathbb{E}\{A_k(t) - R_k(t)|\Theta(t)\} \\
 & \quad + \sum_{n=1}^N H_n(t) \mathbb{E}\{P_n(t) - P_n^{avg}(t)|\Theta(t)\}.
 \end{aligned} \tag{44}$$

where B satisfies (22).

APPENDIX B PROOF OF THEOREM 1

Since we use a C -additive approximation algorithm, which yields a value within a constant C of the infimum of the R.H.S of (21), it is easy to obtain the following:

$$\begin{aligned}
 & \Delta(\Theta(t)) - V\mathbb{E}\{\eta_{EE}(t)|\Theta(t)\} \\
 & \leq B + C - V\mathbb{E}\{\eta_{EE}^*(t)|\Theta(t)\} \\
 & \quad + \sum_{n=1}^N H_n(t) \mathbb{E}\{P_n^*(t) - P_n^{avg}| \Theta(t)\} \\
 & \quad + \sum_{k=1}^{K_R} Q_k(t) \mathbb{E}\{A_k(t) - R_k^*(t)|\Theta(t)\},
 \end{aligned} \tag{45}$$

where $R_k^*(t)$, $P_n^*(t)$ and $\eta_{EE}^*(t)$ are corresponding values for stationary algorithm referred to in Lemma 3. Substituting (37) into (45) and taking the limit as $\theta \rightarrow 0$ leads to:

$$\Delta(\Theta(t)) - V\mathbb{E}\{\eta_{EE}(t)|\Theta(t)\} \leq B + C - V\eta_{EE}^{opt} - \sum_{k=1}^{K_R} \epsilon Q_k(t). \tag{46}$$

Summing (46) over $t \in \{0, 2, \dots, T-1\}$, we obtain

$$\begin{aligned}
 & \mathbb{E}\{L(\Theta(T))\} - \mathbb{E}\{L(\Theta(0))\} - V \sum_{t=0}^{T-1} \mathbb{E}\{\eta_{EE}(t)|\Theta(t)\} \\
 & \leq (B + C)T - VT\eta_{EE}^{opt} - \sum_{t=0}^{T-1} \sum_{k=1}^{K_R} \epsilon Q_k(t).
 \end{aligned} \tag{47}$$

Based on the fact that $Q_k(t) \geq 0$ for all t and the assumption

(35), we rearrange (47) and obtain

$$\begin{aligned}
 & \mathbb{E}\{Q_k^2(T)\} \\
 & \leq 2(B + C - V\eta_{EE}^{opt}) + 2VT\eta_{EE}^{max} + 2\mathbb{E}\{L(\Theta(0))\}.
 \end{aligned} \tag{48}$$

According to the fact that $\{\mathbb{E}\{|Q_k(T)|\}\}^2 \leq \mathbb{E}\{Q_k^2(T)\}$, we have

$$\begin{aligned}
 & \mathbb{E}\{|Q_k(T)|\} \\
 & \leq \sqrt{2T(B + C - V\eta_{EE}^{opt}) + 2VT\eta_{EE}^{max} + 2\mathbb{E}\{L(\Theta(0))\}}.
 \end{aligned} \tag{49}$$

Dividing (49) by T and taking limits as $T \rightarrow \infty$

$$\lim_{T \rightarrow \infty} \frac{\mathbb{E}\{Q_k(T)\}}{T} = 0. \tag{50}$$

Thus, queues are mean-rate stable from Definition 1, which indicates that constraint C2 is satisfied according to the proposed algorithm. Similar proof can be applied to $H_n(t)$.

APPENDIX C PROOF OF THEOREM 2

Based on the inequality (47) obtained in Appendix B and $\mathbb{E}\{L(\Theta(0))\} < \infty$, we obtain

$$V \sum_{t=0}^{T-1} \mathbb{E}\{\eta_{EE}(t)|\Theta(t)\} \geq VT\eta_{EE}^{opt} - (B + C)T - \mathbb{E}\{L(\Theta(0))\}, \tag{51}$$

with some non-negative terms neglected when appropriate.

Dividing both sides of (51) by VT and taking the limit as $T \rightarrow \infty$, we obtain

$$\overline{\eta_{EE}} \geq \eta_{EE}^{opt} - \frac{B + C}{V}. \tag{52}$$

Thus Theorem 2 is proved.

APPENDIX D PROOF OF THEOREM 3

According to the fact that $\mathbb{E}\{L(\Theta(T))\} < \infty$, (47) can be re-written as

$$\begin{aligned}
 & \sum_{t=0}^{T-1} \sum_{k=1}^{K_R} \epsilon Q_k(t) \leq (B + C)T - VT\eta_{EE}^{opt} - \mathbb{E}\{L(\Theta(T))\} \\
 & \quad + \mathbb{E}\{L(\Theta(0))\} + V \sum_{t=0}^{T-1} \mathbb{E}\{\eta_{EE}(t)|\Theta(t)\} \\
 & \leq (B + C)T - VT\eta_{EE}^{opt} - \mathbb{E}\{L(\Theta(T))\} \\
 & \quad + \mathbb{E}\{L(\Theta(0))\} + VT\eta_{EE}^{max}(t).
 \end{aligned} \tag{53}$$

Dividing (53) by ϵT and taking the limit as $T \rightarrow \infty$, the following is obtained:

$$\overline{Q} \leq \frac{B + C + V(\eta_{EE}^{max} - \eta_{EE}^{opt})}{\epsilon}. \tag{54}$$

This completes the proof of Theorem 3.

REFERENCES

- [1] J. Wu, Z. Zhang, Y. Hong, and Y. Wen, "Cloud radio access network (C-RAN): A primer" *IEEE Network*, vol. 29, pp. 35-42, Jan. 2015.
- [2] M. Peng, C. Wang, V. Lau, and H. V. Poor, "Fronthaul-constrained cloud radio access networks: Insights and challenges," *IEEE Wireless Commun.*, vol. 22, no. 2, pp. 152-160, Apr. 2015.

- [3] M. Peng, Y. Li, J. Jiang, J. Li, and C. Wang, "Heterogeneous cloud radio access networks: A new perspective for enhancing spectral and energy efficiencies," *IEEE Wireless Commun.*, vol. 21, no. 6, pp. 126–135, Dec. 2014.
- [4] M. Peng, Y. Li, Z. Zhao, and C. Wang, "System architecture and key technologies for 5G heterogeneous cloud radio access networks," *IEEE Network*, vol. 29, no. 2, pp. 6–14, Mar. 2015.
- [5] M. Peng, H. Xiang, Y. Cheng, S. Yan, and H. Poor, "Inter-tier interference suppression in heterogeneous cloud radio access networks," *IEEE Access*, vol. 2015, pp. xxx-xxx, Dec. 2015.
- [6] Y. Shi, J. Zhang, and K. Lataief, "Group sparse beamforming for green cloud-RAN," *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, pp. 2809–2823, May 2014.
- [7] A. Liu and V. Lau, "Joint power and antenna selection optimization in large cloud radio access network," *IEEE Trans. Signal Process.*, vol. 62, no. 5, pp. 1319–1328, Mar. 2014.
- [8] R. Gupta, E. Strinati, and D. Ktenas, "Energy efficient joint DTX and MIMO in cloud radio access networks," *IEEE CLOUDNET*, Paris, France, pp.191-196, Nov. 2012.
- [9] S. Luo, R. Zhang, and T. Lim, "Downlink and uplink energy minimization through user association and beamforming in cloud RAN," *IEEE Trans. Wireless Commun.*, vol. 14, no. 1, pp. 494–508, Jan. 2015.
- [10] X. Ge *et al.*, "Energy-efficiency optimization for MIMO-OFDM mobile multimedia communication systems with QoS constraints," *IEEE Trans. Veh. Technol.*, vol. 63, no. 5, pp. 2127–2138, Jun. 2014.
- [11] R. Xie, F. Yu, H. Ji, and Y. Li, "Energy-efficient resource allocation for heterogeneous cognitive radio networks with femtocells," *IEEE Trans. Wireless Commun.*, vol. 11, no. 11, pp. 3910–3920, Nov. 2012.
- [12] D. Cao, S. Zhou, and Z. Niu, "Improving the energy efficiency of two-tier heterogeneous cellular networks through partial spectrum reuse," *IEEE Trans. Wireless Commun.* vol. 12, no. 8, pp. 4129–4141, Aug. 2013.
- [13] Z. Xu, *et al.*, "Energy-efficient CoMP precoding in heterogeneous networks," *IEEE Trans. Signal Process.* vol. 62, no. 4, pp. 1005–1017, Feb. 2014.
- [14] L. Zhou and H. Wang, "Toward blind scheduling in mobile media cloud: Fairness, simplicity, and asymptotic optimality," *IEEE Tran. Multimedia*, vol. 15, no. 4, pp.735-746, Jun. 2013.
- [15] F. Meshkati, H. V. Poor, S. C. Schwartz, and R. V. Balan, "Energy-efficient resource allocation in wireless networks with quality-of-service constraints," *IEEE Trans. Commun.*, vol. 57, no. 11, pp. 3406–3414, Nov. 2009.
- [16] F. Meshkati, H. V. Poor, and S. C. Schwartz, "Energy efficiency-delay tradeoffs in CDMA networks: A game-theoretic approach," *IEEE Trans. Inf. Theory*, vol. 55, no. 7, pp. 3220–3228, Jul. 2009.
- [17] C. Lin, M. Tao, G. Stuber, and Y. Liu, "Distributed cross-layer resource allocation for statistical QoS provisioning in femtocell networks," in *Proc. Int. Conf. Commun.*, Budapest, Hungary, pp. 5000–5004, Sep. 2013.
- [18] J. Yang and S. Ulukus, "Trading rate for balanced queue lengths for network delay minimization," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 5, pp. 988–996, May. 2011.
- [19] V. Lau and Y. Cui, "Delay-optimal power and subcarrier allocation for OFDMA systems via stochastic approximation," *IEEE Trans. Wireless Commun.*, vol. 9, no. 1, pp. 227–233, Jan. 2010.
- [20] A. Liu and V. Lau, "Cache-enabled opportunistic cooperative MIMO for video streaming in wireless systems," *IEEE Trans. Signal Process.*, vol. 62, no. 2, pp. 390–402, Jan. 2014.
- [21] H. Ju, B. Liang, J. Li, and X. Yang, "Dynamic joint resource optimization for LTE-Advanced relay networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 11, pp. 5668–5678, Nov. 2013.
- [22] E. Stai and S. Papavasiliou, "User optimal throughput-delay trade-off in multihop networks under NUM framework," *IEEE Commun. Lett.*, vol. 18, no. 11, pp. 1999–2002, Nov. 2014.
- [23] R. Urganonkar and M. Neely, "Opportunistic cooperation in cognitive femtocell networks," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 607–616, Apr. 2012.
- [24] J. Chen and V. Lau, "Large deviation delay analysis of queue-aware multi-user MIMO systems with two-timescale mobile-driven feedback," *IEEE Trans. Signal Process.*, vol. 61, no. 16, pp. 4067–4076, Aug. 2013.
- [25] S. Christensen, R. Agarwal, E. Carvalho, and J. Cioffi, "Weighted sum-rate maximization using weighted MMSE for MIMO-BC beamforming design," *IEEE Trans. Wireless Commun.*, vol. 7, no. 12, pp. 4792–4799, Dec. 2008.
- [26] B. Dai and W. Yu, "Sparse beamforming and user-centric clustering for downlink cloud radio access network," *IEEE Access*, vol. 2, pp. 1326–1339, Dec. 2014.
- [27] M. Joham, W. Utschick, and J. Nosssek, "Linear transmit processing in MIMO communications systems," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 2700–2712, Aug. 2005.
- [28] M. Neely, *Stochastic network optimization with application to communication and queuing systems*. Morgan & Claypool, 2010.
- [29] C. He, B. Sheng, P. Zhu, X. You, and G. Y. Li, "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.
- [30] Z. Zhou, M. Dong, K. Ota, J. Wu, and T. Sato, "Energy efficiency and spectral efficiency tradeoff in device-to-device (D2D) communications" *IEEE Wireless Commun. Lett.*, vol. 3, no. 5, pp. 485–488, Oct. 2014.
- [31] S. H. Park, O. Simeone, O. Sahin, and S. Shamai, "Joint precoding and multivariate backhaul compression for the downlink of cloud radio access networks," *IEEE Trans. Signal Process.*, vol. 61, pp. 5646–5658, Nov. 2013.
- [32] E. Candes, M. Wakin, and S. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5, pp. 877–905, 2008.
- [33] B. Dai and Y. Wei, "Sparse beamforming for limited-backhaul network MIMO system via reweighted power minimization," *IEEE Globecom* Atlanta, USA, Dec. 2013, pp. 1962–1967.
- [34] S. Christensen, R. Agarwal, E. Carvalho, and J. M. Cioffi, "Weighted sumrate maximization using weighted MMSE for MIMO-BC beamforming design," *IEEE Trans. Wireless Commun.*, vol. 7, no. 12, pp. 4792–4799, Dec. 2008.
- [35] Q. Shi, M. Razaviyayn, Z. Luo, and C. He, "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.
- [36] S. Kaviani, O. Simeone, W. Krzymien, and S. Shamai, "Linear precoding and equalization for network MIMO with partial cooperation," *IEEE Trans. Veh. Technol.*, vol. 61, no. 5, pp. 2083–2096, Jun. 2012.
- [37] M. Grant and S. Boyd, "CVX: matlab software for disciplined convex programming, version 2.0 beta," [Online] Available: <http://cvxr.com/cvx>, Sep. 2013.
- [38] H. Shen, B. Li, M. Tao, and X. Wang, "MSE-based transceiver designs for the MIMO interference channel," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3480–3489, Nov. 2010.
- [39] D. Bertsekas and R. Gallager, *Data Networks*, Prentice-Hall, 1987.