Energy Efficient Joint Resource Scheduling for Delay-Aware Traffic in Cloud-RAN

Kaiwei Wang, Wuyang Zhou
Dept.of Electronic Engineering and Information Science
University of Science and Technology of China
Hefei, Anhui, 230027, P.R.China
Email: wangkw@mail.ustc.edu.cn, wyzhou@ustc.edu.cn

Shiwen Mao
Dept. of Electrical & Computer Engineering
Auburn University
Auburn, AL 36849-5201, U.S.A
Email: smao@ieee.org

***Abstract*—In this paper, we focus on the energy efficient joint resource scheduling scheme in time varying Cloud-RAN with delay sensitive traffic. We jointly consider the computation resources provided by baseband units (BBUs), which are modeled as the data processing rate of each virtual machine (VM) provided by the BBUs, and the antenna resources provided by the remote radio heads (RRHs), which are modeled as the beamforming vectors for each user equipment (UE) considering the limited fronthaul capacity and per-UE QoS requirement. Based on the Lyapunov optimization method, we divide the original problem into two subproblems, i.e., the BBU processing rate scheduling problem and the network-wide beamforming strategy problem. The first subproblem can be formulated as a convex programming problem, and we can solve the second subproblem with a weighted minimum mean square error (WMMSE) approach. With the detailed theoretical analysis and simulation results, it is clear that we can achieve a trade-off between energy efficiency and traffic delay, which can be controlled by the control parameter V.**

I. Introduction

To meet the future demand of growing mobile data traffic and high-speed data applications in wireless communication systems, cloud computing technologies, which can provide great flexibility needed for radio access networks, have received considerable attention for the deployment of mobile core network functionalities. As a promising application of the cloud concept, a new cellular architecture named Cloud Radio Access Network (Cloud-RAN) has been proposed by the industry and studied by several researchers. Unlike the existing cellular network, the computation resources for baseband processing in each baseband unit (BBU) are centralized in Cloud-RAN to form a baseband resource pool (BRP). The remote radio heads (RRHs) where the radio function is located, are connected with the BRP through highly reliable optical fibers. All RRHs can dynamically share the baseband resource provided by any BBU in BRP [1]. Such a centralized structure enables several joint processing techniques such as coordinated multipoint transmission (CoMP) and joint beamforming. However, there are still many new challenges to face, such as energy efficiency, effective resource scheduling, the limited fronthaul and backhaul capacity, virtualization of computational resources in the BBU, and so forth.

There has been some interesting work focused on resource allocation and scheduling in the Cloud-RAN. Some existing work is on bandwidth resource allocation among users in OFDM based Cloud-RAN [2], while many other works have addressed CoMP among different RRHs to capitalize the advantages brought about by the centralized architecture of Cloud-RAN [3], [4]. However, these works only concentrate on the antenna resources of RRHs to serve each user equipment (UE) and have overlooked the computation resources scheduling in the BRP. In a cloud based network system like Cloud-RAN, the data processing center provides computation resources as well as consuming a significant amount of power [5]. Therefore the resource and power consumption of BRP should be important considerations in the resource scheduling problems in Cloud-RAN. In [6], the authors proposed a cross-layer resource allocation scheme that jointly considers resources from BBUs and RRH antennas. This work based on a Poisson traffic process, which may not apply to the more complex general mobile data.

In this work, we propose an energy efficient resource scheduling scheme that jointly considers the scheduling of RRH antenna resources and BBU computation resources. Instead of the simple Poisson process, we use the Lyapunov optimization approach to handle the data traffic from the core network to UEs under a very general traffic model. The Lyapunov method can effectively reduce the implementation complexity without loss of generality, because it does not require any prior knowledge of traffic rates or channel states. In order to precisely model the system performance, we formulate the problem as a long term energy efficiency optimization instead of focusing on the instantaneous system performance as many existed works did. We also consider the fronthaul capacity limitation and single UE's QoS requirement in the analysis. With the Lyapunov method, we derive the system energy efficiency and traffic delay trade-off with a control parameter V. To the best of our knowledge, this is the first energy efficiency and traffic delay trade-off that jointly considering the BBU resources and RRH antenna resources in the Cloud-RAN. The simulation results demonstrate the trade-off and show that different levels of energy efficiency can be achieved through adjusting the control parameter.

The rest of this paper is organized as follows: Section II introduces the Cloud-RAN structure and formulates the energy efficient joint resource scheduling problem; Based on

978-1-5090-1328-9/16/$31.00 ©2016 IEEE

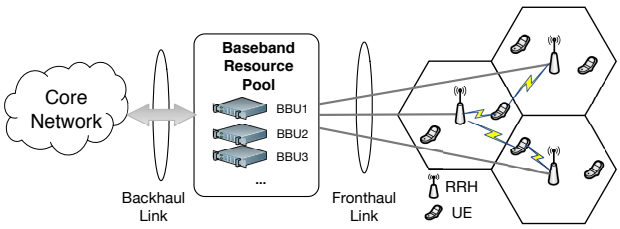

Fig. 1. The Cloud-RAN architecture.

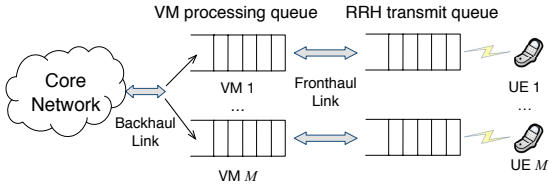

Fig. 2. Queue Model of Cloud-RAN

general Lyapunov optimization approach, Section III gives the theoretical analysis of the problem; Section IV shows the simulation results and finally, Section V provides the conclusion.

II. System Model and Problem Formulation

A. Cloud-RAN with Delay-Aware Traffic

We consider a small cell based Cloud-RAN, which is one of the latest evolutions of cellular networks considering the relatively low transmit power of RRHs, as shown in Fig. 1 [7]. RRHs in different cells can jointly serve all UEs in the area. The RRHs are connected to the BRP with capacity limited fronthaul links. We denote the set of all RRHs as $\mathcal{L} = \{1, 2, \cdots, L\}$. Each RRH is equipped with N antennas. We denote the set of UEs as $\mathcal{M} = \{1, 2, \cdots, M\}$, and each UE is equipped with one antenna. The BRP can serve each UE by generating a virtual machine (VM) to provide computation resources as a common data center does in a cloud based system [8].

The channel gain from RRH l to UE m in time slot t is denoted as $\mathbf{h}_{lm}(t) \in \mathbb{C}^{1\times N}$, and $\mathbf{w}_{lm}(t) \in \mathbb{C}^{N\times 1}$ denotes the beamforming vector for UE m from RRH l. We use $\mathbf{w}_m = \{\mathbf{w}_{jm} | j \in \mathcal{L}\} \in \mathbb{C}^{LN\times 1}$ to denote the beamforming vector of UE m from all RRHs, which also indicates from which RRH each UE receives signals. In $\mathbf{w}_m$, if UE m is not served by RRH j, the beamforming vector $\mathbf{w}_{jm}$ will be set to $\mathbf{0}$. Similarly, $\mathbf{h}_m = \{\mathbf{h}_{jm} | j \in \mathcal{L}\} \in \mathbb{C}^{1\times LN}$ is used to denote the channel gain from all RRHs to UE m. Let x_m represent the data symbol for UE m with $E[|x_m|^2] = 1$ and x_m's be independent to each other.

In this work, we assume that the Cloud-RAN operates in slotted time with slots normalized to integral units. We use t to represent each slot, indicating the time interval $[t, t+1), t \in \{0, 1, 2, \cdots\}$. The received signal at UE m in slot t can then be represented by

$$y_m(t) = \mathbf{h}_m(t)\mathbf{w}_m(t)x_m(t) + \sum_{i\neq m} \mathbf{h}_m(t)\mathbf{w}_i(t)x_i(t) + z_m(t), \quad \forall\, m \in \mathcal{M}, \quad (1)$$

where the first term on the right hand side is the useful signal for UE m, the second term is the interference signal from other active RRHs, and $z_m(t) \sim \mathcal{CN}(0, \sigma_m^2(t))$ is the additive Gaussian noise. The SINR of UE m in slot t is defined as

$$\text{SINR}_m(t) = \frac{|\mathbf{h}_m(t)\mathbf{w}_m(t)|^2}{\sum_{i\neq m} |\mathbf{h}_m(t)\mathbf{w}_i(t)|^2 + \sigma_m^2(t)}, \forall\, m \in \mathcal{M}. \quad (2)$$

Then the nomalized achievable rate (i.e., spectral efficiency) for UE m in time slot t can be formulated as

$$c_m(t) = \log_2(1 + \text{SINR}_m(t)). \quad (3)$$

And the total down-link achievable transmission rate $R_{tot}(t)$ of all UEs in slot t is

$$R_{tot}(t) = \sum_{m\in\mathcal{M}} c_m(t). \quad (4)$$

In this paper, we use a double-layer queueing model to describe the two steps of data transmission from the core network to each UE [6]. As shown in Fig. 2, the UE data from the core network will first come into the BRP and will be assigned to each VM provided by the BBUs. After being processed in the VM processing queue, the data will be transmitted to the RRHs that serve the UE, and then to the UE through the wireless channel.

We use two queues to describe the data processing behavior of the VM and the data transmitting behavior of the RRH, respectively. Let $H_m(t)$ denote the processing queue of VM m in the BBU pool. Each UE's data will be assigned to one VM, and so VM m will process the data to be transmitted to UE m. We use $A_m(t)$ to denote the data arrives from the core network for UE m, which is assumed to be independent and identically distributed (i.i.d.) with arrival rate $\lambda_m = \mathbb{E}\{A_m(t)\}$. With a processing rate of $\mu_m(t)$, the VM in the BBU pool can deliver the data to the RRH that serves UE m. Then the data process in VM m can be expressed as

$$H_m(t+1) = \max[H_m(t) - \mu_m(t), 0] + A_m(t). \quad (5)$$

We use $Q_m(t)$ to denote the transmitting queue for UE m. With the transmission rate given in (3), the traffic queue dynamics for UE m can be expressed as

$$Q_m(t+1) = \max[Q_m(t) - c_m(t), 0] + \mu_m(t). \quad (6)$$

Based on the beamforming vector from each RRH to different UEs, the transmit power consumption of RRH l can be expressed as

$$P_l^R(t) = \sum_{m\in\mathcal{M}} ||\mathbf{w}_{lm}(t)||_2^2 + P^C = \sum_{m\in\mathcal{M}} ||\mathbf{D}_l\mathbf{w}_m(t)||_2^2 + P^C, \quad (7)$$

where $\mathbf{D}_l = \{\mathbf{0}_N^1, \cdots, \mathbf{I}_N^l, \cdots, \mathbf{0}_N^L, l \in \mathcal{L}\}_{N\times LN}$, and P^C represents the circuit and fronthaul link power consumption of each RRH l. Note that an RRH could be shut down when

there is no UE in its serving cell to save energy. We use $P_m^V(t)$ to denote the power consumption of the VM generated by the BRP to serve UE m in time slot t, and have

$$P_m^V(t) = \varphi_m(\mu_m(t)), \tag{8}$$

where $\varphi_m(\mu_m(t))$ is the power consumption of VM m according to the service rate it provides for UE m. The power consumption model of the VM is often assumed to be a convex and increasing function of the processing rate of the VM such as $\varphi_m(\mu_m(t)) = \kappa(\mu_m(t))^\alpha$ [9]. With the power consumption of each VM in (8), the energy consumption of the entire network in time slot t can be formulated as

$$P_{tot}(t) = \sum_{m\in\mathcal{M}} \varphi_m(\mu_m(t)) + \sum_{l\in\mathcal{L}} \sum_{m\in\mathcal{M}} ||\mathbf{D}_l \mathbf{w}_m(t)||_2^2 + |\mathcal{A}|P^C, \tag{9}$$

where $|\mathcal{A}|$ represents the number of active RRHs, which depends on how many RRHs have UEs in its cell to serve.

B. Problem Formulation

Different from most existing energy efficiency definitions of Cloud-RAN [4], [10], we define the energy efficiency η_{EE} as the ratio of the long term aggregated data delivered to the long term energy consumption of the entire network [11], in units of bit/Hz/J. We have

$$\eta_{EE} = \frac{\lim\limits_{t\to\infty} \frac{1}{t} \sum_{\tau=0}^{t-1} E\{R_{tot}(\tau)\}}{\lim\limits_{t\to\infty} \frac{1}{t} \sum_{\tau=0}^{t-1} E\{P_{tot}(\tau)\}} = \frac{\bar{R}_{tot}(\mathbf{w})}{\bar{P}_{tot}(\mathbf{w}, \boldsymbol{\mu})}, \tag{10}$$

where $\mathbf{w} = \{\mathbf{w}_m | m \in \mathcal{M}\}$ and $\boldsymbol{\mu} = \{\boldsymbol{\mu}_m | m \in \mathcal{M}\}$ are the sets of beamforming vectors and VM processing rates for all UEs and VMs, respectively.

The optimization problem can be formulated as

$$\begin{aligned}
\max_{\{\mathbf{w},\boldsymbol{\mu}\}} \quad & \eta_{EE} = \frac{\bar{R}_{tot}(\mathbf{w})}{\bar{P}_{tot}(\mathbf{w}, \boldsymbol{\mu})} \qquad (11)\\
s.t. \quad & \text{C1}: \text{Queues } Q_m(t) \text{ and } H_m(t) \text{ are mean rate stable,} \\
& \qquad \forall m, t \\
& \text{C2}: \lim_{t\to\infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}\{P_l^R(\tau)\} \le P_{ave}^R, \forall\, l \\
& \text{C3}: \sum_{m\in\mathcal{M}} ||\mathbf{D}_l \mathbf{w}_m(t)||_2^2 \le P_{max}^R, \forall\, l, t \\
& \text{C4}: \sum_{m\in\mathcal{M}} \{||\mathbf{D}_l \mathbf{w}_m(t)||_2^2\} c_m(t) \le R_l, \forall\, l, t \\
& \text{C5}: B_0 \log_2(1 + \text{SINR}_m(t)) \ge r_m, \forall\, m, t.
\end{aligned}$$

In (11), C1 is the queue stability constraint which ensures that all arriving data leaves the queue in a finite time. C2 is the average transmit power constraint of each RRH. C3 is the peak transmit power constraint of each RRH in each time slot. Among these two constraints, P_{ave}^R and P_{max}^R are the average transmit power threshold and maximum transmitting power of each RRH, respectively. C4 is the fronthaul capacity constraint and C5 guarantees the QoS of each UE is satisfied, where B_0 is the total system bandwidth and r_m is the data rate requirement for UE m.

With the concept that a non-convex ℓ_0-norm optimization objective can be approximated by a convex ℓ_1 reweighted norm [12], we can approximate the front-haul capacity constraint C4 as

$$\text{C4}: \sum_{m\in\mathcal{M}} \beta_{lm}(t) \tilde{c}_m(t) ||\mathbf{D}_l \mathbf{w}_m(t)||_2^2 \le R_l, \forall\, l, t, \tag{12}$$

where $\beta_{lm}(t)$ is a constraint weight, which is updated iteratively according to

$$\beta_{lm}(t) = \frac{1}{||\mathbf{D}_l \mathbf{w}_m(t)||_2^2 + \tau}, \ \forall\, l, m, t, \tag{13}$$

and $\tilde{c}_m(t)$ is the optimal RRH transmission rate for UE m obtained by the previous iteration.

Since the phase of $\mathbf{w}_m$ has no impact on the optimization problem and will not change the constraints, we assume that each term of $\mathbf{h}_m(t)\mathbf{w}_m(t)$ has a zero imaginary part. Then we can rewrite the constraint C5 as

$$\text{C5}: \sqrt{\sum_{i\neq m} |\mathbf{h}_m(t)\mathbf{w}_i(t)|^2 + \sigma_m^2} \le \frac{1}{\sqrt{\gamma_m}} \Re(\mathbf{h}_m(t)\mathbf{w}_m(t)), \forall m, t \tag{14}$$

which is a second order cone (SOC) constraint, where $\gamma_m = 2^{\frac{r_m}{B_0}} - 1$ is the equivalent SINR threshold for UE m.

It is obvious that the objective function of (11) is non-convex. Thus it can be classified as non-linear fractional programming, and the maximize energy efficiency η_{EE}^* from the long-term perspective can be defined as

$$\eta_{EE}^* = \frac{\bar{R}_{tot}(\mathbf{w}^*)}{\bar{P}_{tot}(\mathbf{w}^*, \boldsymbol{\mu}^*)}, \tag{15}$$

where $\mathbf{w}^*$ and $\boldsymbol{\mu}^*$ are the optimal beamforming vectors and VM data processing rate, respectively.

Theorem 1. *The maximum energy efficiency η_{EE}^* can be achieved if and only if*

$$\begin{aligned}
&\max \quad \bar{R}_{tot}(\boldsymbol{w}) - \eta_{EE}^* \bar{P}_{tot}(\boldsymbol{w}, \boldsymbol{\mu}) \qquad (16)\\
&= \bar{R}_{tot}(\boldsymbol{w}^*) - \eta_{EE}^* \bar{P}_{tot}(\boldsymbol{w}^*, \boldsymbol{\mu}^*) = 0, \qquad (17)
\end{aligned}$$

where $\boldsymbol{w}$ and $\boldsymbol{\mu}$ are any feasible solutions to (11) with constraints C1-C5.

Proof: The proof is similar to that in Appendix A of [11], and is omitted for brevity.

With Theorem 1, the energy efficiency maximization problem (11) can be equivalently transformed to

$$\begin{aligned}
\max_{\mathbf{w},\boldsymbol{\mu}} \quad & \bar{R}_{tot} - \eta_{EE}^* \bar{P}_{tot} \qquad (18)\\
s.t. \quad & \text{C1} \sim \text{C5}.
\end{aligned}$$

However, the problem is still hard to solve since the optimal value of η_{EE}^* is not known in advance. So we define $\eta_{EE}(t)$ for $t \in 1, 2, \cdots$, with $\eta_{EE}(0) = 0$ as

$$\eta_{EE}(t) = \frac{\sum_{\tau=0}^{t-1} R_{tot}(\tau)}{\sum_{\tau=0}^{t-1} P_{tot}(\tau)}. \tag{19}$$

Replacing η_{EE}^* with $\eta_{EE}(t)$, the problem can be reformulated as

$$\begin{aligned} \max_{\mathbf{w},\boldsymbol{\mu}} \quad & \bar{R}_{tot}(t) - \eta_{EE}(t)\bar{P}_{tot}(t) \\ s.t. \quad & \text{C1} \sim \text{C5}. \end{aligned} \tag{20}$$

We can use an iterative method to obtain the long term optimal solution. This method is proven to be an effective way to solve the problem [11], and it has been widely used to solve many stochastic optimization problems with renewal systems [13].

III. Delay-aware Joint Resource Scheduling Strategy

In this section, we first describe the problem using the Lyapunov method. We then partition the problem into two subproblems, i.e., the BBU processing rate scheduling problem and the network-wide beamforming strategy problem, and analyze them respectively. We propose an iterative algorithm to achieve a long term optimal joint resource scheduling strategy based on the optimization problem (20).

A. General Lyapunov Optimization Approach

Based on the Lyapunov method, we can define virtual power queues to transform the average power constraint C2 into a queue stability problem [14], [15]. The virtual power queue of RRH l can be defined as

$$Y_l(t+1) = \max[Y_l(t) - P_{ave}^R + P_l^R(t), 0]. \tag{21}$$

We denote $\boldsymbol{\Theta}(t) = [\mathbf{H}(t), \mathbf{Q}(t), \mathbf{Y}(t)]$ as the combined matrix of the traffic queues and virtual power queues. We define the Lyapunov function as

$$L(\boldsymbol{\Theta}(t)) = \frac{1}{2}\sum_{m\in\mathcal{M}} H_m(t)^2 + \frac{1}{2}\sum_{m\in\mathcal{M}} Q_m(t)^2 + \frac{1}{2}\sum_{l\in\mathcal{L}} Y_l(t)^2. \tag{22}$$

The penalty in this paper is defined as the energy efficiency, then the Lyapunov drift-plus-penalty on each slot t is formulated as,

$$\begin{aligned} &\Delta L(\boldsymbol{\Theta}(t)) - V\eta_{EE}(t) \\ =&L(\boldsymbol{\Theta}(t+1)) - L(\boldsymbol{\Theta}(t)) + V(\eta_{EE}(t)\bar{P}_{tot}(t) - \bar{R}_{tot}(t)) \end{aligned} \tag{23}$$

and an upper bound of the penalty can be derived as

$$\begin{aligned} &\mathbb{E}\{\Delta L(\boldsymbol{\Theta}(t)|\boldsymbol{\Theta}(t))\} + V\mathbb{E}\{\eta_{EE}(t)P_{tot}(t) - R_{tot}(t)|\boldsymbol{\Theta}(t)\} \\ \leq& B + \mathbb{E}\left\{\sum_{m\in\mathcal{M}}(Q_m(t) - H_m(t))\mu_m(t) + \sum_{m\in\mathcal{M}} H_m(t)A_m(t) + \right. \\ &\left.\sum_{m\in\mathcal{M}} V\eta_{EE}(t)P_m^V(t)|\boldsymbol{\Theta}(t)\right\} + \\ &\mathbb{E}\left\{\sum_{l\in\mathcal{L}}(Y_l(t) + V\eta_{EE}(t))P_l^R(t) - \sum_{l\in\mathcal{L}} Y_l(t)P_{ave}^R - \right. \\ &\left.\sum_{m\in\mathcal{M}}(Q_m(t) + V)c_m(t)|\boldsymbol{\Theta}(t)\right\}. \end{aligned} \tag{24}$$

where B is a positive constant and the proof of it can be found in [11] (omitted for brevity). Base on the general Lyapunov optimization method, our target of an energy efficient joint resource scheduling strategy has become the problem to minimize the right hand side of (24). It is clear that the objective value depends on variables μ_m and c_m, respectively. So we can divide the minimization problem into two subproblems, i.e., the BBU processing control problem and the beamforming strategy problem. Both problems will be analyzed in the following section of this section.

B. BBU Processing Control

For each VM, the BRP needs to decide an optimal processing rate based on the arriving data for each UE and the queue occupancy at the current time slot. According to (24), we can decide the processing rate for each VM separately. The problem could be formulated as

$$\min_{\boldsymbol{\mu}} \quad (Q_m(t) - H_m(t))\mu_m(t) + V\eta_{EE}(t)P_m^V(t) \tag{25}$$

$$s.t. \quad \text{Queues } H_m(t) \text{ is mean rate stable.} \tag{26}$$

Since $P_m^V(t)$ is a convex and increasing function of $\mu_m(t)$, the problem is a convex optimization problem and can be easily solved with a standard convex solver such as CVX [16].

C. Beamforming Strategy

Based on (24) and leaving out the fixed part of the circuit and link power, the beamforming strategy to solve the user scheduling problem can be expressed as

$$\begin{aligned} \min_{\mathbf{w}} \quad & \sum_{l\in\mathcal{L}}(Y_l(t) + V\eta_{EE}(t))\sum_{m\in\mathcal{M}}||\mathbf{D}_l\mathbf{w}_m(t)||_2^2 - \\ & \sum_{m\in\mathcal{M}}(Q_m(t) + V)c_m(t) \\ s.t. \quad & \text{C3} \sim \text{C5}. \end{aligned} \tag{27}$$

Since the problem is still non-convex, we can reformulate it into an equivalent WMMSE problem as [3], which is

$$\begin{aligned} \min_{\mathbf{w}} \quad & \sum_{m\in\mathcal{M}}(Q_m(t) + V)(\omega_m(t)e_m(t) - \log\omega_m(t)) + \\ & \sum_{l\in\mathcal{L}}(Y_l(t) + V\eta_{EE}(t))\sum_{m\in\mathcal{M}}||\mathbf{D}_l\mathbf{w}_m(t)||_2^2 \\ s.t. \quad & \text{C3} \sim \text{C5}, \end{aligned} \tag{28}$$

where $\omega_m(t)$ denote the MSE weight for UE m in time slot t and $e_m(t)$ is the corresponding MSE defined as

$$\begin{aligned} e_m(t) &\triangleq \mathbb{E}\{(s_m(t) - u_m(t)y_m(t))(s_m(t)^* - (u_m(t)y_m(t))^*)\} \\ &= u_m^H(t)\left(\sum_{i\in\mathcal{M}}\mathbf{h}_m(t)\mathbf{w}_i(t)\mathbf{w}_i^H(t)\mathbf{h}_m^H(t) + \sigma^2\mathbf{I}\right)u_m(t) - \\ &\quad 2\Re\{u_m^H(t)\mathbf{h}_m(t)\mathbf{w}_m(t)\} + 1, \end{aligned} \tag{29}$$

under the independent assumption of $s_m(t)$ and $z_m(t)$. The optimum MSE weight $\omega_m(t)$ under the given $\mathbf{w}_{lm}(t)$ and $u_m^H(t)$ can be derived as

$$\omega_m(t) = e_m(t)^{-1}. \tag{30}$$

The optimal receiver under the given $\mathbf{w}_{lm}(t)$ can be derived as

$$u_m(t) = \frac{\mathbf{h}_m(t)\mathbf{w}_m(t)}{\sum_{i\in\mathcal{M}} \mathbf{h}_m(t)\mathbf{w}_i(t)\mathbf{w}_i^H(t)\mathbf{h}_m^H(t) + \sigma^2}. \quad (31)$$

The optimal beamforming vector could then be obtained by solving the following optimization problem.

$$\begin{aligned}\min_{\mathbf{w}} \sum_{m\in\mathcal{M}} \mathbf{w}_m^H(t)\Big(\sum_{i\in\mathcal{M}} (Q_i(t)+V)\omega_i(t)u_i^H(t)(\mathbf{h}_i^{\mathcal{L}_m})^H(t) \\ \mathbf{h}_i^{\mathcal{L}_m}(t)u_i(t) + \sum_{l\in\mathcal{L}_m} (Y_l(t)+V\eta_{EE}(t))\mathbf{D}_l^H\mathbf{D}_l\Big)\mathbf{w}_m(t) - \\ 2\sum_{m\in\mathcal{M}} (Q_m(t)+V)\omega_m(t)\Re\{u_m^H(t)\mathbf{h}_m(t)\mathbf{w}_m(t)\} \quad (32) \\ s.t. \quad \text{C3} \sim \text{C5},\end{aligned}$$

which is a quadratic programming with quadratic constraints and SOC constraints. The problem can be easily solved using a standard convex solver such as CVX [16]. The complete procedure to solve the energy efficient beamforming optimization problem (27) is elaborated in Algorithm 1.

Algorithm 1 Energy Efficient Beamforming Strategy with QoS Constraint

Initialize: Choose network-wide beamforming vector $\mathbf{w}_m(t)$, and based on $\mathbf{w}_m(t)$, initialize $\beta_{lm}(t)$, $\tilde{c}_m(t)$, $\forall\, l, m$.

1: **repeat**
2: With fixed beamforming vector $\mathbf{w}_m(t)$, calculate the MSE weight $\omega_m(t)$ and the optimum receiver $u_m(t)$ according to (29), (30) and (31) in turn, $\forall\, m$;
3: Solve problem (32) with fixed $\omega_m(t)$ and $u_m(t)$ to obtain the optimized energy efficient beamforming vector $\mathbf{w}_m^*$;
4: Compute the UE data rate $c_m(t)$ according to (2) and (3);
5: Update $\tilde{c}_m(t) = c_m(t)$, $\mathbf{w}_m(t) = \mathbf{w}_m^*$, and update $\beta_{lm}(t)$ according to (13), $\forall\, l, m$;
6: **until** convergence

D. Long Term Energy Efficiency Joint Resource Scheduling Algorithm

Based on all the derivations introduced above, the long term energy efficiency joint resource scheduling algorithm is summarized in Algorithm 2, which solves problem (11).

IV. Simulation Results

In this section, we evaluate the system performance of our proposed joint resource scheduling through numerical simulations. The Cloud-RAN consists of seven small cells, and each cell is equipped with one RRH. The simulation parameters are presented in Table I. The number of UEs in each cell is 2. The arrival data for the UEs in each time slot t is assumed to be uniformly distributed in $[0, 2\lambda]$. Without loss of generality, we assume all UEs have the same data arrival rate, i.e., $\lambda_1 = \cdots = \lambda_M = \lambda$. We also have $r_1 = \cdots = r_m = r$ to ensure the basic QoS requirement for each UE.

Algorithm 2 Long Term Energy Efficiency Joint Resource Scheduling Algorithm

1: In each time slot t, observe the queue state $H_m(t)$, $Q_m(t)$, $Y_l(t)$, and $Z_m(t)$, $\eta_{EE}(t)$ and the channel state $\mathbf{h}_m(t), \forall m$;
2: Obtain the optimal VM data processing rate $\mu_m(t)$ by solving the BBU processing control problem (25), $\forall\, m$;
3: Obtain the optimal network wide beamforming vector $\mathbf{w}_m(t)$ with Algorithm 1 and calculate $c_m(t)$;
4: Based on the optimal $\mathbf{w}_m(t)$, calculate RRH power consumption $P_l^R(t)$ according to (7), and then update $H_m(t)$, $Q_m(t)$ and $Y_l(t)$ as in (5), (6), and (21). Calculate $\eta_{EE}(t+1)$ according to (19);

TABLE I
Simulation Parameters

Parameter	Value
System Bandwidth	10MHz
Distance between each RRH	100m
Distance-dependent Path Loss From RRH to UE	$140.7 + 36.7 * \lg(R)$, R in kilometers
Log-normal Shadowing	-8dB
Penetration Loss	-20dB
Maximum RRH Transmit Power P_{max}^R	30dBm
Number of antennas of each RRH N	2
Noise Power Density	-174dBm/Hz

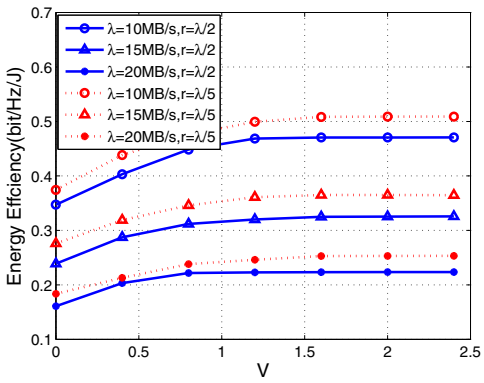

Fig. 3. Energy efficiency against control parameter V

Fig. 3 illustrates the relationship between the control parameter V and the system level energy efficiency η_{EE}. We can see clearly that the energy efficiency increases first and then converge to a steady value as V is increased. While with increasing data arrival rate λ, η_{EE} obviously decreases, due to the fact that both BBU and RRH need to consume more power, which offsets the gains achieved by a higher data rate. We also simulate the scenario where the UEs have different QoS requirements, and we can see that the energy efficiency becomes higher when we reduce the QoS requirement of the UEs, due to the same reason when the data arrival rate is low.

Fig. 4 shows that the average queue length grows linearly with the increased control parameter V (i.e., as $O(V)$). Furthermore, the queue length also increases based on the arrival

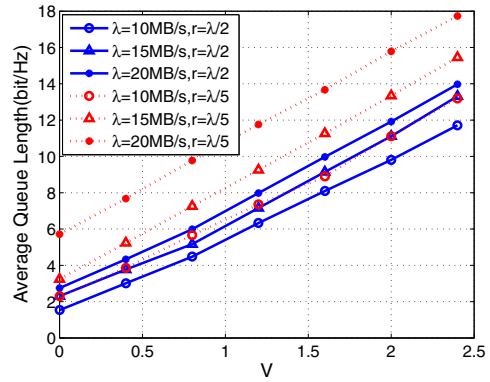

Fig. 4. Average queue length against control parameter V

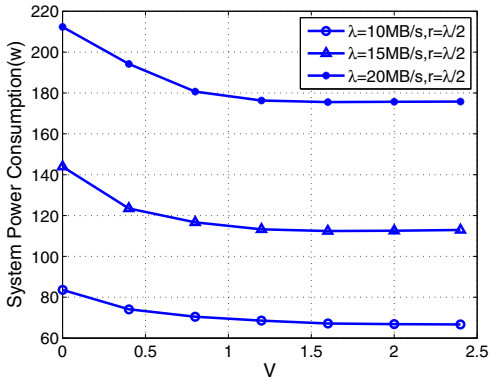

Fig. 5. System power consumption against control parameter V

data rate λ. Combined with Fig. 3, we can see the trade-off between system level energy efficiency and data queue backlog (indicative of UE's delay performance), which can be useful in the design of Cloud-RAN. When we need a higher energy efficiency, we can use a larger V. On the contrary, when the UEs have a tighter delay requirement, a smaller V may be needed. When we lower the QoS requirements, the queue backlog becomes longer, because there is less data to be processed and transmitted in each time slot.

Fig. 5 expresses the relationship of varying control parameter V against the total system power consumption. The system consumes more power as the data arrival rate is increased. With a larger V, the total energy consumption decreases, because the system emphasizes more on energy efficiency with a larger V. This leads to a relatively lower processing rate for the VM in the BBU. Therefore the processing power of BBU obviously decreases, and then the entire system energy consumption decreases.

V. Conslusions

In this paper, we proposed an energy efficient joint resource scheduling scheme in Cloud-RAN considering both the computation resources in BBUs and the antenna resources provided by RRHs. Using a Lyapunov optimization approach, we partitioned the problem as two subproblems. The first subproblem can be easily solved with a convex solver. For the second subproblem, we used a WMMSE approach to obtain network-wide energy efficient beamforming vectors, taking the limited fronthaul capacity and per-UE QoS requirements into consideration. To precisely describe the system performance, we proposed a long term energy efficient optimization algorithm instead of focusing on the instantaneous system performance. Simulation results demonstrated the trade-off between energy efficiency and delay as it was analyzed by the theoretical derivation.

Acknowledgment

The work of K. Wang and W. Zhou has been supported by the Natural Science Foundation of China under Grant 61461136002, the Fundamental Research Funds for the Central Universities under Grant WK3500000003, Huawei Technology Innovative Research; and the work of S. Mao has been supported by the U.S. National Science Foundation under Grant CNS-1247955 and the Wireless Engineering Research and Education Center at Auburn University.

References

[1] China Mobile Res. Inst., "C-RAN: the road towards green RAN," White Paper, ver. 2.5, Beijing, China, Oct. 2011.

[2] M. Peng, K. Zhang, J. Jiang, J. Wang, and W. Wang, "Energy-efficient resource assignment and power allocation in heterogeneous cloud radio access networks," *IEEE Trans. Veh. Technol.*, vol.64, no.11, pp.5275–5287, Nov. 2015.

[3] B. Dai and W. Yu, "Sparse beamforming and user-centric clustering for downlink cloud radio access network," *IEEE Access J.*, vol.2, pp.1326–1339, Oct. 2014.

[4] X. Huang, G. Xue, R. Yu, and S. Leng, "Joint scheduling and beamforming coordination in cloud radio access networks with QoS guarantees," *IEEE Trans. Veh. Technol.*, vol.PP, no.99, pp.1, Aug. 2015.

[5] A. Greenberg, J. Hamilton, D. A. Maltz, and P. Patel. "The cost of a cloud: Research problems in data center networks," *ACM SIGCOMM Computer Communication Review*, vol.39, no.1, pp.68–73, Jan. 2009.

[6] J. Tang, W.P. Tay, and T.Q.S. Quek, "Cross-layer resource allocation with elastic service scaling in cloud radio access network," *IEEE Trans. Wireless Commun.*, vol.14, no.9, pp.5068–5081, Sept. 2015.

[7] W. Ni and I.B. Collings, "A new adaptive small-cell architecture," *IEEE J. Sel. Areas Commun.*, vol.31, no.5, pp.829–839, May 2013.

[8] R. Urgaonkar, U. Kozat, K. Igarashi, and M. Neely, "Dynamic resource allocation and power management in virtualized data centers," in *Proc. IEEE/IFIP NOMS'10*, Osaka, Japan, Apr. 2010, pp.479–486.

[9] W. Zhang, Y. Wen, K. Guan, D. Kilper, H. Luo, and D.O. Wu, "Energy-optimal mobile cloud computing under stochastic wireless channel," *IEEE Trans. Wireless Commun.*, vol.12, no.9, pp.4569–4581, Sept. 2013.

[10] H. Xiang, Y. Yu, Z. Zhao, Y. Li, and M. Peng, "Tradeoff between energy efficiency and queues delay in heterogeneous cloud radio access networks," in *Proc. IEEE ICC Workshops*, London, UK, June 2015, pp.2727–2731.

[11] Y. Li, M. Sheng, Y. Shi, X. Ma, and W. Jiao, "Energy efficiency and delay tradeoff for time-varying and interference-free wireless networks," *IEEE Trans. Wireless Commun.*, vol.13, no.11, pp.5921–5931, Nov. 2014.

[12] E. Candes, M. Wakin, and S. Boyd,"Enhancing sparsity by reweighted ℓ_1 minimization," *J. Fourier Analysis Appl.*, vol.14, no.5, pp.877–905, Dec. 2008.

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

[14] Z. Jiang and S. Mao, "Energy delay trade-off in cloud offloading for multi-core mobile devices," *IEEE Access J.*, vol.3, no.1, pp.2306–2316, Nov. 2015.

[15] Y. Huang, S. Mao, and R.M. Nelms, "Adaptive electricity scheduling in microgrids," *IEEE Trans. Smart Grid*, vol.5, no.1, pp.270–281, Jan. 2014.

[16] M. Grant and S. Boyd. *CVX: Matlab Software for Disciplined Convex Programming*, Version 2.0 Beta, Sept. 2013. [Online]. Available: http://cvxr.com/cvx