Title	Cross-layer resource allocation with elastic service scaling in cloud radio access network
Author(s)	Tang, Jianhua; Tay, Wee Peng; Quek, Tony Q. S.
Citation	Tang, J., Tay, W. P., & Quek, T. Q. S. (2015). Cross-layer resource allocation with elastic service scaling in cloud radio access network. IEEE Transactions on Wireless Communications, 14(9), 5068-5081. doi:10.1109/TWC.2015.2432023
Date	2015
URL	http://hdl.handle.net/10220/47838
Rights	© 2015 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works. The published version is available at: https://doi.org/10.1109/TWC.2015.2432023.

Cross-Layer Resource Allocation with Elastic Service Scaling in Cloud Radio Access Network

Jianhua Tang, *Student Member, IEEE*, Wee Peng Tay, *Senior Member, IEEE*, and Tony Q.S. Quek, *Senior Member, IEEE*

***Abstract*—Cloud radio access network (C-RAN) aims to improve spectrum and energy efficiency of wireless networks by migrating conventional distributed base station functionalities into a centralized cloud baseband unit (BBU) pool. We propose and investigate a cross-layer resource allocation model for C-RAN to minimize the overall system power consumption in the BBU pool, fiber links and the remote radio heads (RRHs). We characterize the cross-layer resource allocation problem as a mixed-integer nonlinear programming (MINLP), which jointly considers elastic service scaling, RRH selection, and joint beamforming. The MINLP is however a combinatorial optimization problem and NP-hard. We relax the original MINLP problem into an extended sum-utility maximization (ESUM) problem, and propose two different solution approaches. We also propose a low-complexity Shaping-and-Pruning (SP) algorithm to obtain a sparse solution for the active RRH set. Simulation results suggest that the average sparsity of the solution given by our SP algorithm is close to that obtained by a recently proposed greedy selection algorithm, which has higher computational complexity. Furthermore, our proposed cross-layer resource allocation is more energy efficient than the greedy selection and successive selection algorithms.**

***Index Terms*—C-RAN, elastic service scaling, cross-layer design, green communication, weighted sum-rate maximization**

I. Introduction

Cloud radio access network (C-RAN) has emerged as a promising solution to the operation and bandwidth challenges faced by future mobile communication infrastructures, which are required to handle an exponentially increasing demand for data traffic [1]. A C-RAN utilizes centralized signal processing in the baseband unit (BBU) pool instead of processing at distributed base stations (BSs), which can result in significant capital and operating expenditure savings. Centralized processing at the BBU pool also allows cooperation between multiple remote radio heads (RRHs), thus improving spectrum efficiency and link reliability. Furthermore, the use of cloud computing technologies as the infrastructure of the BBU pool greatly improves hardware utilization.

Decoupling the baseband signal processing from the RRHs is the most attractive feature of C-RAN, which means that RRHs only need to keep the basic transmission and reception functionalities, while computationally intensive tasks can be migrated to the BBU pool in a cloud data center. This centralized signal processing and scheduling feature in the BBU pool further makes a variety of prospective technologies feasible, including centralized encoding and decoding, centralized compression and decompression, and joint beamforming.

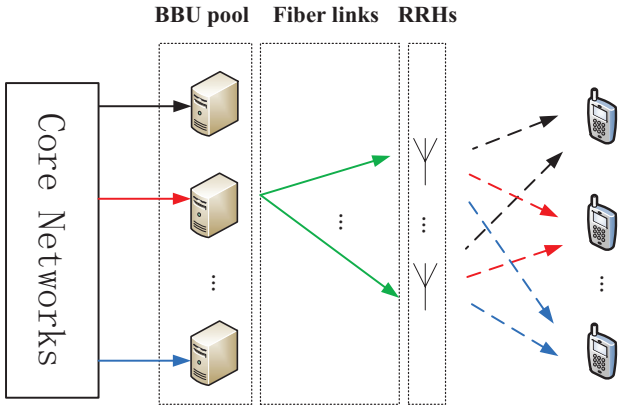


Fig. 1. Cloud radio access network (C-RAN).

Although C-RAN makes it possible to transition conventional cellular networks (CCNs) from hardware defined infrastructures to a software defined environment, many design and operational challenges that have been resolved in CCNs need to be revisited in C-RAN. One particular example of importance is the resource allocation problem. Specifically, in CCNs, power control and beamforming strategies have been used to minimize the system power consumption such that users' predefined quality-of-service (QoS) requirements are fulfilled. Unfortunately, these strategies cannot plug directly into the C-RAN framework. In CCNs, the BSs' computation capacity is fixed. As a result, resource allocation methods in CCNs are oblivious to the computation capacities of the BSs although users' achievable QoS levels are actually dependent on them. Under the C-RAN architecture, the computational functionalities in conventional BSs are migrated to the cloud based virtual machines (VMs) in the BBU pool, whose computation capacity can be scaled according to users' QoS requirements and various parameters from different layers of the OSI stack, including the incoming traffic rate from the application layer and wireless channel state information from the physical layer. Therefore, developing a cross-layer resource allocation scheme is required in order to fully utilize the features of a C-RAN, and to optimize the overall system power consumption.

Most of the power in a C-RAN are consumed in the

J. Tang and W. P. Tay are with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. e-mail: {jtang4, wptay}@ntu.edu.sg.

T. Q. S. Quek is with Singapore University of Technology and Design. He is also with Institute for Infocomm Research, A*STAR, Singapore. e-mail: tonyquek@sutd.edu.sg.

following subsytems:

- **BBU pool**. Having a cloud data center as the BBU pool not only allows centralized signal processing, but also elastic service scaling of the resources of the BBU pool. Specifically, the BBU pool can dynamically adjust each VM capacity to optimize power consumption for changing traffic and channel states. An important research challenge is to design low-complexity algorithms for dynamic service scaling in the BBU pool.
- **Fiber links**. Each RRH is connected to the BBU pool via a high-bandwidth, low-latency fiber link. The power consumption in the fronthaul links has traditionally been ignored in the CCN literature since it is relatively much lower than the power consumption in the BSs of the CCN. However, in C-RAN, the power consumption in fronthaul links is comparable to the power consumption at the RRHs since the RRHs are architecturally much simpler compared to conventional BSs. By turning off some redundant fiber links, as well as the RRH connected to these fiber links, energy savings can be achieved. This motivates the link or RRH selection problem.
- **RRHs**. In C-RAN, the functionality of RRHs can be as simple as just a signal transmission and reception point. The RRHs can cooperate with each other to perform centralized joint beamforming to mitigate interference. Thus, the throughput of the wireless channels to the users can be significantly enhanced. An important research issue here is the design of the joint beamforming at the RRHs in order to achieve an optimal trade-off in channel throughput and energy efficiency.

In this paper, we consider the problem of optimizing the allocated VM computation capacities in the BBU pool, the set of selected RRHs, and the beamforming strategies at the active RRHs in order to minimize the overall system power consumption for C-RAN.

A. Related work

C-RAN aims to be a competitive and potential 5G framework, which has been attracting comprehensive research attention from both industry and academia since 2011 [2]–[5]. Many prototypes, test-beds and architecture designs have been done to show the feasibility and performance gain by adopting C-RAN [6]–[9]. The concept of RAN as a service (RANaaS) has also been developed based on the structure of C-RAN [10].

In the fronthaul of C-RAN, i.e., the RRHs and the wireless channel, Coordinated Multipoint (CoMP) techniques are deployed to enhance the system throughput. In order to enhance energy efficiency [11], cell, BS or RRH selection for the fronthaul has been comprehensively studied over the past several years [12]–[17]. For example, the authors in [12] and [14] jointly consider the base station selection problem and linear precoding. Fronthaul link or RRH selection for C-RAN have been studied by [18]–[23]. For instance, [19] considers joint BS selection and distributed compression in C-RAN to improve energy efficiency, while [22] considers RRH selection jointly with fronthaul beamforming to minimize the system power consumption.

The joint RRH selection and beamforming problem is NP-hard [22], therefore to solve it exactly is computationally intractable when the number of RRHs is large. We summarize some commonly used approaches here. In the first approach, the problem is formulated as a mixed-integer nonlinear programming (MINLP), and then solved by Branch and Bound (BnB) or Branch and Cut (BnC) methods [17]. Both the BnB and BnC methods yield the optimal solution, but have high time complexity. Another approach is the "*sorting-and-removing*" method [13], [22], in which the RRHs are ranked according to some priority criteria in each iteration, and the RRH with the lowest priority is removed. The process continues until the problem becomes infeasible. This method can produce a near-optimal solution, but still has high computational complexity. The "*sparsity-inducing*" method is inspired by compressive sensing. Reweighted l_1-norm relaxation and sparsity-inducing norms are used to obtain a sparse subset of RRHs [19], [20], [22]. This method is efficient in computational complexity but cannot guarantee optimality. Finally, constructing a Markov Chain Monte Carlo (MCMC) is a potential way to solve the RRH selection problem as well [16]. In this work, we propose a "*Shaping-and-Pruning*" method, which is a trade-off between the sorting-and-removing and sparsity-inducing methods, in order to obtain a near-optimal performance with lower computational complexity.

The BBU pool of C-RAN comprises many general purpose processors (GPP), which forms a cloud computing infrastructure using virtualization technology [24]. A computationally aware strategy is proposed to reduce the computational outage in C-RAN recently [25]. However, most of the previous works related to the C-RAN BBU pool just makes use of the centralized processing property offered by cloud computing to optimize the system. For example, a central encoder is developed in [26] to jointly encode the messages intended for the mobile stations, and a cloud decoder in [19] utilizes the joint statistics of the received correlated signals to decompress the received signal. Theses works do not consider the issue of elastic service scaling and resource allocation the BBU pool, which is one of the focus of this paper. In addition, unlike most of the works in the literature, which investigate methods to provide QoS guarantees for specific layers in the OSI stack (e.g., ensuring bandwidth or latency requirements are met only in the wireless transmission part), we consider methods to ensure cross-layer QoS guarantees in this paper.

B. Main contributions

In this paper, we formulate the cross-layer resource allocation problem as a MINLP by minimizing the system power consumption, which consists of three parts: the power consumption in the BBU pool with respect to (w.r.t.) the VM computation capacity, the power consumption in the fiber fronthaul links w.r.t. the number of links (or, active RRHs) and the transmission power on the RRHs w.r.t. the transmit beamformer. We relax the MINLP into an extended sum-utility maximization (ESUM) problem, and propose two different approximate solution approaches. In the first approach, we approximate the ESUM problem as a quasi weighted sum-rate maximization (QWSRM) problem, and propose a BnB

algorithm to solve it. The QWSRM problem is an extension of the weighted sum-rate maximization (WSRM) problem, which has been studied in [27], [28]. In the second approach, we utilize the weighted minimum mean square error (WMMSE) method to obtain a locally optimal solution to the ESUM problem. Based on the achievable rates found by either solving the QWSRM problem or using the WMMSE approach, we propose an efficient Shaping-and-Pruning algorithm to perform RRH selection. Our proposed algorithm achieves a trade-off between computational complexity and solution optimality. We provide simulation results that suggest that our proposed approach outperforms the recently proposed greedy selection algorithm of [22] and successive selection algorithm of [29] in terms of overall system power consumption, since these methods only optimize the RRH selection and RRH beamforming strategies. This shows that cross-layer optimization can result in higher energy efficiencies for a C-RAN.

The remainder of this paper is organized as follows. We present the C-RAN system model in Section II, and introduce the QWSRM problem and its solution in Section III. In Section IV, we formulate the minimization of system power consumption as a MINLP, approximate this MINLP problem as a QWSRM problem, and propose an efficient algorithm to solve it. We present simulation results in Section V, and conclude the paper in Section VI.

Notations. We use boldface lower case letters to denote vectors. The notation $\|\mathbf{x}\|_2$ is the Euclidean norm of $\mathbf{x}$, while $(\cdot)^T$ and $(\cdot)^H$ represent transpose and conjugate transpose, respectively. We use $\mathbb{C}$ to denote the set of complex numbers, and $\mathcal{CN}(0,\sigma^2)$ to denote the distribution of a circularly symmetric complex normal zero mean random variable with variance σ^2. The $\log$ function is the logarithm function with base 2.

II. System model

In this section, we present our C-RAN system model and problem formulation. Suppose that there are N single-antenna user equipments (UEs) and L available RRHs, each with K antennas, in a C-RAN cluster. We denote the sets of all UEs and all RRHs as $\mathcal{N} = \{1,\cdots,N\}$ and $\mathcal{L} = \{1,\cdots,L\}$, respectively. We denote the set of active RRHs (i.e., the set of RRHs that are servicing the UEs in $\mathcal{N}$, and their associated fiber links) as $\mathcal{A}$. We have $\mathcal{A} \subseteq \mathcal{L}$. The amount of voice and data traffic associated with each UE $i \in \mathcal{N}$ up to time t is given by $\Delta_i(t)$ (bits), and each UE i is served by one VM with computation capacity μ_i in the BBU pool. After processing by the VM, the data is forwarded to the UE via $|\mathcal{A}|$ active RRHs (we assume data sharing among the RRHs), where $|\mathcal{A}| \leq L$ is the cardinality of the set $\mathcal{A}$. Let the achievable wireless transmission rate to UE i using the active RRHs be c_i.

Queueing model, with the channel capacity as the queue's service rate, is widely used to characterize wireless communication systems [30]. Therefore, we introduce a double-layer queueing network to represent each UE's data processing and transmitting behavior in the C-RAN downlink (cf. Figure 2). Our model can be easily extended to the C-RAN uplink as well. Specifically, in the BBU pool, the data of UE i is processed (e.g., encoded) by a VM, which is abstracted as a processing queue, with mean service rate μ_i. Then, the processed data is transmitted to UE i via the RRHs, which are modeled using a transmitting queue with mean service rate c_i. Note that the links between the BBU pool and the RRHs are high-bandwidth, low-latency optical fiber links with negligible transmission delay. However, the power consumption P_f of each fiber link cannot be neglected, compared with the power consumption in the associated RRH.

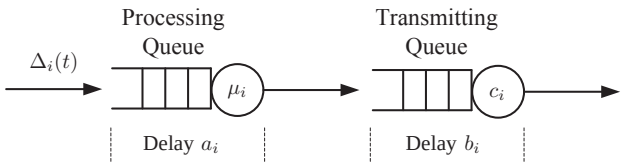


Fig. 2. Queueing model representation of a C-RAN processing and transmission path for a UE i.

For each UE $i \in \mathcal{N}$, let a_i represent the expected delay in the processing queue (i.e., the expected delay in the BBU pool) and b_i be the expected delay in the transmitting queue (i.e., the expected delay during wireless transmission). Our goal is to design a cross-layer algorithm such that for each UE i, the system expected delay $d_i = a_i + b_i$ satisfies the cross-layer QoS constraint:

$$d_i \leq \tau_i, \tag{1}$$

where τ_i is a predefined QoS requirement for UE i.

We assume that UE i's packet arrival process to the processing queue is a Poisson process with mean rate $\lambda_i > 0$, where $\Delta_i(t) = \lambda_i t$, and the service time of each data packet in the processing queue follows an exponential distribution with mean $1/\mu_i$. Then, the arrival process to the transmitting queue is the same as the one to the processing queue [31], [32]. We assume that the service time of each data packet in the transmitting queue follows an exponential distribution with mean $1/c_i$. Therefore, the data processing and transmitting for each UE i in our C-RAN model can be treated as two M/M/1 queues [33] in tandem. We have for $\mu_i, c_i > \lambda_i$,

$$d_i = \frac{1}{\mu_i - \lambda_i} + \frac{1}{c_i - \lambda_i}.$$

In the wireless transmission, C-RAN leverages CoMP transmission to enhance the throughput [34]. There are two types of CoMP techniques in the downlink: coordinated scheduling/coordinated beamforming (CS/CB) and joint transmission (JT). In this work, we consider JT as the CoMP technique in C-RAN, i.e., each UE's data can be shared among all the coordinated RRHs. Let x_i denote the data symbol for the ith UE with $E[|x_i|^2] = 1$, and $\mathbf{w}_{ij} \in \mathbb{C}^K$ denote the transmit beamformer for the UE i from RRH j. The channel from RRH j to UE i is denoted as $\mathbf{h}_{ij}^H$, where $\mathbf{h}_{ij} \in \mathbb{C}^K$, for $i \in \mathcal{N}$ and $j \in \mathcal{A}$. Thus, the received signal at UE i is given by

$$\hat{x}_i = \sum_{j\in\mathcal{A}} \mathbf{h}_{ij}^H \mathbf{w}_{ij} x_i + \sum_{k\neq i}^{N} \sum_{j\in\mathcal{A}} \mathbf{h}_{ij}^H \mathbf{w}_{kj} x_k + \delta_i,$$

where the first term on the right hand side is the useful signal

for UE i, the second term is the interference to UE i, and $\delta_i \sim \mathcal{CN}(0, \sigma_i^2)$ is the additive white Gaussian noise (AWGN) at UE i.

As such, the signal-to-interference-plus-noise ratio (SINR) at UE i, with the active RRH set $\mathcal{A}$, becomes

$$\text{SINR}_i(\mathcal{A}) = \frac{|\sum_{j\in\mathcal{A}} \mathbf{h}_{ij}^H \mathbf{w}_{ij}|^2}{\sigma_i^2 + \sum_{k\neq i}^{N} |\sum_{j\in\mathcal{A}} \mathbf{h}_{ij}^H \mathbf{w}_{kj}|^2}. \tag{2}$$

The achievable rate of UE i, c_i, should satisfy

$$c_i \leq B_i \log(1 + \text{SINR}_i(\mathcal{A})), \tag{3}$$

where B_i is the bandwidth for UE i. Each RRH j has maximum transmitting power constraint given by

$$\sum_{i=1}^{N} \mathbf{w}_{ij}^H \mathbf{w}_{ij} = \sum_{i=1}^{N} \|\mathbf{w}_{ij}\|_2^2 \leq E_j, \text{ for } j \in \mathcal{L}. \tag{4}$$

A. Problem formulation

The BBU pool of C-RAN utilizes a cloud computing infrastructure with elastic service scaling. In particular, the BBU pool can dynamically adjust the VMs' computation capacities to handle dynamic user traffic and channel states. We model VM i's power consumption $\varphi_i(\mu_i)$ as a function of its computation capacity μ_i. We make the following assumptions regarding the VM's power consumption function φ_i.

Assumption 1: For each VM i, $i \in \mathcal{N}$, the power consumption function $\varphi_i(\mu_i)$ has the following properties:

1) $\varphi_i(\mu_i) \geq 0$ for all $\mu_i \geq 0$,
2) $\varphi_i(\mu_i)$ is a convex and increasing function of μ_i.

The power consumption of a VM i is often modeled as $\varphi_i(\mu_i) = k_i \mu_i^{a_i}$, where $k_i > 0$ and $a_i > 1$ are positive constants. This power consumption function satisfies Assumption 1, and has been widely adopted in the literature [35]–[38].

Our aim is to minimize the system power consumption in C-RAN, which consists of three components: the power consumption in the BBU pool, the power consumption in the fiber links, and the power consumption at the RRHs. Specifically, (i) the power consumption for each VM in the BBU pool with computation capacity μ_i is $\varphi_i(\mu_i)$, $\forall i \in \mathcal{N}$; (ii) the power consumption for each active fiber link is P_f; and (iii) the power consumption at RRH $j \in \mathcal{A}$ is $(1/\eta)\sum_{i=1}^{N} \mathbf{w}_{ij}^H \mathbf{w}_{ij}$, where $\eta \in (0, 1)$ is the inefficiency coefficient of the amplifier in each RRH. Our optimization problem can then be formulated as follows:

$$\text{(P0)} \quad \min_{\mu_i, c_i, \mathbf{w}_{ij}, \mathcal{A}} \quad \sum_{i=1}^{N} \varphi_i(\mu_i) + |\mathcal{A}| P_f + \frac{1}{\eta} \sum_{i=1}^{N} \sum_{j\in\mathcal{A}} \mathbf{w}_{ij}^H \mathbf{w}_{ij}$$

$$\text{s.t.} \quad \frac{1}{\mu_i - \lambda_i} + \frac{1}{c_i - \lambda_i} \leq \tau_i, \ \forall i \in \mathcal{N}, \tag{5}$$

$$\lambda_i < \mu_i, \lambda_i < c_i, \ \forall i \in \mathcal{N}, \tag{6}$$

$$c_i \leq B_i \log\left(1 + \text{SINR}_i(\mathcal{A})\right), \ \forall i \in \mathcal{N}, \tag{7}$$

$$\sum_{i=1}^{N} \mathbf{w}_{ij}^H \mathbf{w}_{ij} \leq E_j, \ \forall j \in \mathcal{L}, \tag{8}$$

where $\text{SINR}_i(\mathcal{A})$ is given by (2), and "s.t." stands for "subject to".

Problem (P0) is difficult to solve for the following reasons: (i) it is a combinatorial optimization problem and NP-hard [20]; and (ii) the problem is nonconvex even if the active RRH set $\mathcal{A}$ is known a priori. However, by first relaxing (P0) into a ESUM problem, and then into a QWSRM problem, we obtain a BnB solution. In the following section, we first discuss the QWSRM problem and its BnB solution.

III. The QWSRM problem

In this section, we extend the classical WSRM problem to the QWSRM problem, and then propose a BnB solution for the latter. The QWSRM problem will be used in Section IV to tackle problem (P0). Throughout this section, we assume that the active RRH set $\mathcal{A}$ is known.

Mathematically, the WSRM problem is typically formulated as

$$\min_{c_i, \mathbf{w}_{ij}} \quad \sum_{i=1}^{N} -\varepsilon_i c_i \tag{9}$$

$$\text{s.t.} \quad c_i \leq B_i \log\left(1 + \text{SINR}_i(\mathcal{A})\right), \ \forall i \in \mathcal{N}, \tag{10}$$

$$\sum_{i=1}^{N} \mathbf{w}_{ij}^H \mathbf{w}_{ij} \leq E_j, \ \forall j \in \mathcal{A}, \tag{11}$$

where c_i is the throughput of UE i, ε_i is an arbitrary nonnegative weight, and $\text{SINR}_i(\mathcal{A})$ is given by (2).

Since the phase rotation of the complex vector $\mathbf{w}_{ij}$ has no impact on the WSRM problem, we can recast the constraint (10) as

$$\|\mathbf{r}_i(\mathcal{A})\|_2 \leq \sqrt{1 + 1/(2^{c_i/B_i} - 1)}\, \mathfrak{Re}[R_{ii}(\mathcal{A})], \ \forall i \in \mathcal{N}, \tag{12}$$

where vector $\mathbf{r}_i(\mathcal{A}) = [R_{i1}(\mathcal{A}), \cdots, R_{iN}(\mathcal{A}), \sigma_i]^T$, $R_{ik}(\mathcal{A}) = \sum_{j\in\mathcal{A}} \mathbf{h}_{ij}^H \mathbf{w}_{kj}$, and $\mathfrak{Re}(\cdot)$ stands for the real part of a complex number [29], [39]. Note that the constraint (12) is a second-order cone (SOC) constraint only if c_i is a constant.

Applying the Cauchy-Schwarz inequality to (7), we have

$$\begin{aligned} c_i &\leq B_i \log\left(1 + \frac{1}{\sigma_i^2} \sum_{j\in\mathcal{A}} \|\mathbf{h}_{ij}\|_2^2 \sum_{j\in\mathcal{A}} \|\mathbf{w}_{ij}\|_2^2\right) \\ &\leq B_i \log\left(1 + \frac{1}{\sigma_i^2} \sum_{j\in\mathcal{A}} \|\mathbf{h}_{ij}\|_2^2 E_j\right) \triangleq \bar{c}_i. \end{aligned} \tag{13}$$

Let $\bar{\mathbf{c}} = [\bar{c}_1, \cdots, \bar{c}_N]^T$.

We define a generalization of the WSRM problem, which has the same constraints as the WSRM problem but with an extended objective function as follows:

$$\min_{c_i, \mathbf{w}_{ij}} \quad f(\mathbf{c}) \tag{14}$$

$$\text{s.t.} \quad \text{(11) and (12)},$$

where $\mathbf{c} = [c_1, \cdots, c_N]^T$, and the objective function $f(\mathbf{c})$, for $0 \leq \mathbf{c} \leq \bar{\mathbf{c}}$, has the following properties:

1) $f(\mathbf{c})$ is a function only of $\mathbf{c}$ and

2) $f(\mathbf{c}) < \infty$ is continuously differentiable, and
3) $f(\mathbf{c})$ is convex in the feasible region defined by (11) and (12).

To avoid trivial solutions for (14), in what follows, we assume $\frac{\partial f(\mathbf{c})}{\partial c_i}\big|_{\mathbf{c}=0} < 0$, for $i = 1 \cdots N$. We call (14) the QWSRM problem, which models a variety of problems in wireless communications. For example, we can interpret $f(\mathbf{c})$ as the *reverse utility* function, corresponding to the concave utility function in network congestion control problems [40], [41].

Let $\tilde{\mathbf{c}} = [\tilde{c}_1, \ldots, \tilde{c}_N]^T$ be the root to the system of equations $\partial f(\mathbf{c})/\partial c_i = 0$, for all $i \in \mathcal{N}$, where each $\tilde{c}_i$ is set to $\bar{c}_i$ if the solution does not exist. Let $\mathcal{F}$ represent the feasible region of the variable $\mathbf{c}$ in (14), and $\mathbf{c}^* = [c_1^*, \ldots, c_N^*]^T$ and $\mathbf{w}_{ij}^*$, $\forall i \in \mathcal{N}, \forall j \in \mathcal{A}$, be the optimal solution of the QWSRM problem.

Theorem 1: The optimal achievable rate $\mathbf{c}^*$ of the QWSRM problem falls inside or on the surface of the N-dimensional rectangle $\mathcal{Q}_{\text{init}} = \{\mathbf{c} \mid c_i \in [0, \min\{\tilde{c}_i,\ \bar{c}_i\}], i \in \mathcal{N}\}$.

Proof: See Appendix A. ■

A WMMSE approach to solve the WSRM problem based on the relationship between SINR and MMSE is proposed in [28]. However, the WMMSE approach cannot always find the global optimal solution. Subsequently, a BnB method is proposed in [27], which shows that this method can produce the global optimal solution. The proposed BnB method uses the fact that the objective function in WSRM problem is monotonically non-increasing in the achievable rates $c_i \geq 0$, for all $i \in \mathcal{N}$. In what follows, we first give a brief overview of the BnB algorithm from [27], and then show how to extend it to solve the QWSRM problem in (14).

The BnB approach is widely used in nonconvex optimization problems, e.g., the integer programming problems. For each iteration step of the BnB algorithm, one needs to generate a sequence of asymptotically tight upper and lower bounds for the objective function, with both bounds converging to the global optimal value eventually. The basic idea in [27] of using the BnB algorithm to solve the WSRM problem is to first expand the unknown feasible region of the WSRM problem to a known initial N-dimensional rectangle, and then sequentially shrink the rectangle until it is small enough, where at each iteration, the variables $\mathbf{c}$ are fixed, and a feasibility problem w.r.t. to the variables $\{\mathbf{w}_{ij} : i \in \mathcal{N}, j \in \mathcal{A}\}$ is solved. This avoids having to solve the nonconvex WSRM problem w.r.t. $\{c_i, \mathbf{w}_{ij}\}$ directly.

Inspired by the BnB algorithm in [27], we develop a similar BnB procedure in Algorithm 1 for the QWSRM problem (14). We use $\mathcal{Q}_{\text{init}}$ given in Theorem 1 as the initial N-dimensional rectangle. We shrink the N-dimensional rectangle by making use of the following upper bound

$$\gamma_{\text{ub}}(\mathcal{Q}) = \begin{cases} f(\mathbf{c}_{\min}), & \mathbf{c}_{\min} \in \mathcal{F} \\ +\infty, & \text{otherwise} \end{cases} \tag{15}$$

and lower bound [1]

$$\gamma_{\text{lb}}(\mathcal{Q}) = \begin{cases} f(\mathbf{c}_{\max}), & \mathbf{c}_{\min} \in \mathcal{F} \\ +\infty, & \text{otherwise,} \end{cases} \tag{16}$$

for every N-dimensional rectangle $\mathcal{Q} \triangleq \{\mathbf{c} \mid c_{i,\min} \leq c_i \leq c_{i,\max}, \forall i \in \mathcal{N}\} \subseteq \mathcal{Q}_{\text{init}}$, where $c_{i,\min}$ and $c_{i,\max}$ denotes the end points of the ith edge of $\mathcal{Q}$, $\mathbf{c}_{\min} = [c_{1,\min}, \cdots, c_{N,\min}]^T$ and $\mathbf{c}_{\max} = [c_{1,\max}, \cdots, c_{N,\max}]^T$. Note that $\mathbf{c}_{\max}$ need not be in the feasible region $\mathcal{F}$. At each iteration, for a given $\mathcal{Q}$, the following feasibility problem is solved:

$$\begin{aligned} \text{find} \quad & \mathbf{w}_{ij}, \forall i \in \mathcal{N},\ \forall j \in \mathcal{A} \\ \text{s.t.} \quad & \|\mathbf{r}_i(\mathcal{A})\|_2 \leq \sqrt{1 + \frac{1}{2^{c_{i,\min}/B_i} - 1}}\, \mathfrak{Re}[R_{ii}(\mathcal{A})],\ \forall i \in \mathcal{N}, \\ & \sum_{i=1}^{N} \mathbf{w}_{ij}^H \mathbf{w}_{ij} \leq E_j,\ \forall j \in \mathcal{A}. \end{aligned} \tag{17}$$

Note that (17) is a second-order cone programming (SOCP) feasibility problem w.r.t. $\mathbf{w}_{ij}$, which can be solved by using interior-point methods on an equivalent SOCP with a trivial objective function [42].

Algorithm 1 BnB algorithm for QWSRM problem

1: Input: $\mathcal{Q}_{\text{init}}$, $\mathcal{A}$, and $f(\mathbf{c})$.
2: Initialize: Obtain $\tilde{c}_i$ by solving $\frac{\partial f(\mathbf{c})}{\partial c_i} = 0$, for $i \in \mathcal{N}$. Set $k = 1$, $\mathcal{B}_1 = \mathcal{Q}_{\text{init}}$, $u_1 = \gamma_{\text{ub}}(\mathcal{Q}_{\text{init}})$ and $l_1 = \gamma_{\text{lb}}(\mathcal{Q}_{\text{init}})$.
3: Check the feasibility of problem (17) with given $\tilde{\mathbf{c}}$.
4: **if** feasible **then**
5: $\mathbf{c}_o = \tilde{\mathbf{c}}$;
6: **else**
7: **while** $u_k - l_k > \epsilon$ **do**
8: Branching:
- Set $\mathcal{Q}_k = \mathcal{Q}$, where $\mathcal{Q}$ satisfies $\gamma_{\text{lb}}(\mathcal{Q}) = l_k$.
- Split $\mathcal{Q}$ into $\mathcal{Q}_I$ and $\mathcal{Q}_{II}$, along one of its longest edges.
- Update $\mathcal{B}_{k+1} = (\mathcal{B}_k \backslash \{\mathcal{Q}_k\}) \cup (\mathcal{Q}_I, \mathcal{Q}_{II})$.

9: Bounding:
- Update $u_{k+1} = \min_{\mathcal{Q} \in \mathcal{B}_{k+1}} \{\gamma_{\text{ub}}(\mathcal{Q})\}$.
- Update $l_{k+1} = \min_{\mathcal{Q} \in \mathcal{B}_{k+1}} \{\gamma_{\text{lb}}(\mathcal{Q})\}$.

10: Set $k = k + 1$;
11: **end while**
12: Set $\mathbf{c}_o = \mathbf{c}_{\min}$;
13: **end if**
14: Output: $\mathbf{c}_o$.

The rationale of using the BnB algorithm to solve the QWSRM problem is the same as that for the WSRM problem, i.e., we sequentially shrink the given N-dimensional rectangle $\mathcal{Q}_{\text{init}}$, where the optimal solution falls in, until the the lower and upper bounds satisfy $u_k - l_k \leq \epsilon$, where $\epsilon > 0$ is a predefined accuracy level. The following result shows that Algorithm 1 converges to the optimal solution of the QWSRM. The proof is similar to Theorem 1 in [27] and the convergence analysis in [43], which we omit for brevity.

[1] Although [27] provides an improved lower bound with additional computational overhead, we use the basic lower bound in this paper for simplicity.

Theorem 2: The output $\mathbf{c}_o$ generated by Algorithm 1, converges arbitrarily close to the optimal solution $\mathbf{c}^*$ of the QWSRM problem, within a finite number of iterations, i.e., for any $\epsilon > 0$, there exists $M > 0$ such that $u_M - f(\mathbf{c}^*) \leq \epsilon$.

Remark 1: The reason that the upper and lower bounds in (15) and (16) are suitable for the QWSRM problem is $f(\mathbf{c})$ is monotonic in each interval $c_i \in [0, \tilde{c}_i]$, for all $i \in \mathcal{N}$. Thus, by fixing the variable $\mathbf{c}$ in the QWSRM problem, instead of solving the nonconvex problem (14) directly, we just need to solve a SOCP (17) in each iteration of the BnB algorithm. For Algorithm 1, the input N-dimensional rectangle $\mathcal{Q}_{\text{init}}$ provided by Theorem 1 can be further shrunk if a priori upper and lower bounds of $\mathbf{c}$ are known.

IV. Cross-layer power consumption minimization

In this section, we reformulate problem (P0) into a MINLP (P1), which we further decompose into a ESUM problem and a RRH selection problem.

In fact, RRH j is inactive means that there is no signal transmitted from RRH j to all the UEs. Hence, RRH j is active or not is equivalent to $\sum_{i=1}^{N} \|\mathbf{w}_{ij}\|_2^2 > 0$ or $= 0$, respectively. In addition, for $i \in \mathcal{N}$,

$$\sum_{j\in\mathcal{A}} \mathbf{w}_{ij}^H \mathbf{w}_{ij} = \sum_{j=1}^{L} \mathbf{w}_{ij}^H \mathbf{w}_{ij}, \tag{18}$$

since $\sum_{j\in\mathcal{A}^c} \mathbf{w}_{ij}^H \mathbf{w}_{ij} = 0$, where $\mathcal{A}^c$ is the complementary set of $\mathcal{A}$.

Let the vector

$$\mathbf{m} = \big[\sum_{i=1}^{N} \|w_{i1}\|_2^2, \sum_{i=1}^{N} \|w_{i2}\|_2^2, \cdots, \sum_{i=1}^{N} \|w_{iL}\|_2^2 \big]^T.$$

Hence, $|\mathcal{A}| = \|\mathbf{m}\|_0$. Combining (18), (12) and $|\mathcal{A}| = \|\mathbf{m}\|_0$, we can reformulate problem (P0) as

$$\begin{aligned}
\text{(P1)} \quad \min_{\mu_i, c_i, \mathbf{w}_{ij}} \quad & \sum_{i=1}^{N} \varphi_i(\mu_i) + \|\mathbf{m}\|_0 P_f + \frac{1}{\eta} \sum_{i=1}^{N}\sum_{j=1}^{L} \mathbf{w}_{ij}^H \mathbf{w}_{ij} \\
\text{s.t.} \quad & \frac{1}{\mu_i - \lambda_i} + \frac{1}{c_i - \lambda_i} \leq \tau_i, \ \forall i \in \mathcal{N}, \\
& \lambda_i < \mu_i, \lambda_i < c_i, \ \forall i \in \mathcal{N}, \\
& \|\mathbf{r}_i(\mathcal{L})\|_2 \leq \sqrt{1 + \frac{1}{2^{c_i/B_i} - 1}} \mathfrak{Re}[R_{ii}(\mathcal{L})], \\
& \qquad \forall i \in \mathcal{N}, \qquad (19) \\
& \sum_{i=1}^{N} \mathbf{w}_{ij}^H \mathbf{w}_{ij} \leq E_j, \ \forall j \in \mathcal{L},
\end{aligned}$$

where (19) is derived from (12) due to the fact that $\mathbf{w}_{ij} = 0$, for $j \in \mathcal{A}^c$.

Proposition 1: In problem (P1), constraint (5) is an active inequality constraint, i.e., the optimal $\{\mu_i, c_i\}$ for problem (P1) satisfies the following equation:

$$\mu_i = \lambda_i + \frac{1}{\tau_i} + \frac{1}{\tau_i^2 (c_i - \lambda_i) - \tau_i}, \ \forall \ i \in \mathcal{N}.$$

Proof: See Appendix B. ∎

Based on Proposition 1, we let

$$g_i(c_i) \triangleq \varphi_i(\mu_i) = \varphi_i \left(\lambda_i + \frac{1}{\tau_i} + \frac{1}{\tau_i^2 (c_i - \lambda_i) - \tau_i} \right), \tag{20}$$

where

$$c_i > \lambda_i + \frac{1}{\tau_i}, \forall i \in \mathcal{N}. \tag{21}$$

Since problem (P1) is a MINLP, we now propose a two-step approach to find an approximate solution to it (cf. Figure 3). Specifically, we first relax problem (P1) to the ESUM problem (P2) below. Then, for problem (P2), we propose two different algorithms to solve it in Section IV-A and Section IV-B respectively. Using the optimal achievable rates obtained by solving problem (P2), problem (P1) becomes a RRH selection problem (cf. problem (Q2) in Section IV-C). We finally propose an efficient Shaping-and-Pruning algorithm to obtain the sparse solution of the RRH selection problem.

Inspired by compressive sensing, l_1-norm is utilized as a convex relaxation of l_0-norm, since the l_1-norm is a convex envelop of the l_0-norm [20]. We can apply l_1-norm relaxation to the objective function of problem (P1) and using Proposition 1, we obtain the following ESUM problem, which is a nonconvex optimization problem.

$$\begin{aligned}
\text{(P2)} \quad \min_{c_i, \mathbf{w}_{ij}} \quad & \sum_{i=1}^{N} g_i(c_i) + (P_f + \frac{1}{\eta}) \sum_{i=1}^{N}\sum_{j=1}^{L} \mathbf{w}_{ij}^H \mathbf{w}_{ij} \\
\text{s.t.} \quad & \|\mathbf{r}_i(\mathcal{L})\|_2 \leq \sqrt{1 + \frac{1}{2^{c_i/B_i} - 1}} \mathfrak{Re}[R_{ii}(\mathcal{L})], \\
& \qquad \forall i \in \mathcal{N}, \\
& c_i > \lambda_i + \frac{1}{\tau_i}, \forall i \in \mathcal{N}. \\
& \sum_{i=1}^{N} \mathbf{w}_{ij}^H \mathbf{w}_{ij} \leq E_j, \ \forall j \in \mathcal{L}.
\end{aligned}$$

A. Approximating the ESUM problem with a QWSRM problem

Similar to (13), we apply the Cauchy-Schwarz inequality to (7), and combining with (18), we have

$$c_i \leq B_i \log \left(1 + \frac{1}{\sigma_i^2} \sum_{j=1}^{L} \|\mathbf{h}_{ij}\|_2^2 \sum_{j=1}^{L} \|\mathbf{w}_{ij}\|_2^2 \right), \forall i \in \mathcal{N}, \tag{22}$$

which further yields

$$\sum_{j=1}^{L} \mathbf{w}_{ij}^H \mathbf{w}_{ij} \geq \frac{(2^{c_i/B_i} - 1)\sigma_i^2}{\sum_{j=1}^{L} \|\mathbf{h}_{ij}\|_2^2}. \tag{23}$$

Hence, problem (P2) can be approximated as

$$\begin{aligned}
\text{(Q1-1)} \quad \min_{c_i, \mathbf{w}_{ij}} \quad & \sum_{i=1}^{N} f_i(c_i) \\
\text{s.t.} \quad & \text{constraints in problem (P2)},
\end{aligned}$$

where

$$f_i(c_i) = g_i(c_i) + (P_f + \frac{1}{\eta}) \frac{(2^{c_i/B_i} - 1)\sigma_i^2}{\sum_{j=1}^{L} \|\mathbf{h}_{ij}\|_2^2}.$$

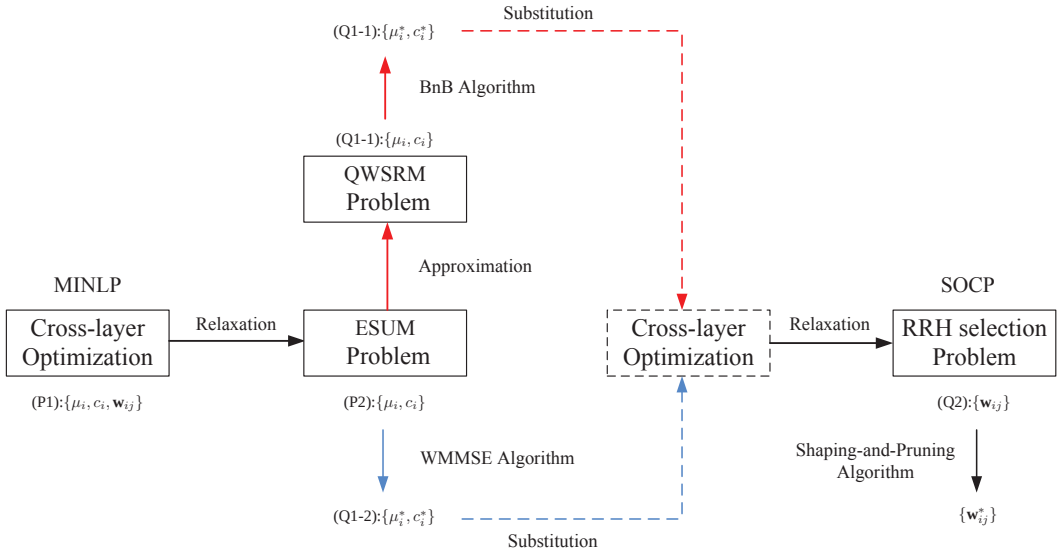


Fig. 3. The two-step approach to solve problem (P1).

Proposition 2: Suppose that Assumption 1 holds. Then, problem (Q1-1) is a QWSRM problem, whose optimal solution $\mathbf{c}^*$ falls inside or on the surface of the N-dimensional rectangle $\hat{\mathcal{Q}}_{\text{init}} = \{\mathbf{c} = [c_1, \ldots, c_N]^T \mid c_i \in (\lambda_i + 1/\tau_i, \min\{\tilde{c}_i, \bar{c}_i\}]$, for $i \in \mathcal{N}\}$, where $\tilde{c}_i$ is the root of the equation $\partial f_i(c_i)/\partial c_i = 0$, and $\bar{c}_i = B_i \log\left(1 + \frac{1}{\sigma_i^2}\sum_{j=1}^{L} \|\mathbf{h}_{ij}\|_2^2 E_j\right)$.

Proof: See Appendix C. ∎

To obtain the optimal achievable rates $\mathbf{c}^* = [c_1^*, \ldots, c_N^*]^T$ for (Q1-1), we utilize Algorithm 1 with the following inputs:

1) $f(\mathbf{c}) = \sum_{i=1}^{N} f_i(c_i)$, where $\mathbf{c} = [c_1, \ldots, c_N]^T$
2) $\mathcal{Q}_{\text{init}} = \hat{\mathcal{Q}}_{\text{init}}$
3) $\mathcal{A} = \mathcal{L}$

The optimal VM computation capacity for UE i as implied by (Q1-1) is then given by $\mu_i^* = \lambda_i + \frac{1}{\tau_i} + \frac{1}{\tau_i^2(c_i^* - \lambda_i) - \tau_i}$. Note that the solution $\{(\mu_i^*, c_i^*) \mid i \in \mathcal{N}\}$ is in general sub-optimal for (P1) because of the relaxations we have done to obtain (Q1-1), but is guaranteed to be feasible for (P1).

B. A WMMSE approach to solve the ESUM problem

Although problem (P2) can be approximated as a QWSRM problem (Q1-1) and then solved by the BnB algorithm, which obtains the global optimal solution for problem (Q1-1), the complexity of the BnB algorithm is still high. In this subsection, we develop a lower complexity algorithm to obtain a local optimal solution for problem (P2) directly. This is done based on an extension of the algorithm proposed by [28], which aims to solve the sum-utility maximization problem using the WMMSE algorithm.

Let $\theta_i(\cdot) = g_i(-B_i \log(\cdot))$ and denote $\tilde{\theta}_i(\cdot)$ as the inverse mapping of the gradient map $\nabla\theta_i(e_i)$. It can be verified that $\theta_i(e_i)$ is a strictly concave function in the interval $(2^{-\hat{c}_i}, \infty)$, where $\hat{c}_i = (\lambda_i + 1/\tau_i)/B_i$.

Proposition 3: The optimal transmit beamformer vectors $\mathbf{w}_{ij}$ for the ESUM problem are the beamforming solutions of the following problem:

$$\text{(Q1-2)} \quad \min_{x_i, y_i, \mathbf{w}_{ij}} \quad \sum_{i=1}^{N} x_i e_i + \sum_{i=1}^{N} \theta_i(\tilde{\theta}_i(x_i)) - \sum_{i=1}^{N} x_i \tilde{\theta}_i(x_i) + (P_f + \frac{1}{\eta}) \sum_{i=1}^{N}\sum_{j=1}^{L} \mathbf{w}_{ij}^H \mathbf{w}_{ij}$$

$$\text{s.t.} \quad \|\mathbf{r}_i\|_2 < \sqrt{1 + 1/(2^{\hat{c}_i} - 1)}\,\Re\mathfrak{e}[R_{ii}], \ \forall i \in \mathcal{N},$$

$$\sum_{i=1}^{N} \mathbf{w}_{ij}^H \mathbf{w}_{ij} \le E_j, \ \forall j \in \mathcal{L},$$

where $x_i > 0$ is the MSE weight for UE i, y_i is the receive beamformer weight at UE i (note that each UE has a single antenna), and

$$\begin{aligned} e_i &\triangleq \mathbb{E}\left[\left\|y_i^H \hat{u}_i - u_i\right\|_2\right] \\ &= \left|y_i^H \sum_{j\in\mathcal{L}} \mathbf{h}_{ij}^H \mathbf{w}_{ij} - 1\right|^2 + \sum_{l\neq i}^{N} \left|y_i^H \sum_{j\in\mathcal{L}} \mathbf{h}_{ij}^H \mathbf{w}_{lj}\right|^2 + \sigma_i^2 |y_i|^2, \\ &= \sum_{l=1}^{N} \left|y_i^H \sum_{j\in\mathcal{L}} \mathbf{h}_{ij}^H \mathbf{w}_{lj}\right|^2 - 2\Re\mathfrak{e}\left[y_i^H \sum_{j\in\mathcal{L}} \mathbf{h}_{ij}^H \mathbf{w}_{ij}\right] + \sigma_i^2 |y_i|^2 + 1. \end{aligned} \tag{24}$$

Moreover, $\theta_i(\tilde{\theta}_i(x_i)) - x_i\tilde{\theta}_i(x_i)$ is strictly convex w.r.t. x_i.

Proof: The proof is similar to that in Appendix B of [28], and is omitted for brevity. ∎

From Proposition 3, instead of solving problem (P2) directly, we can solve problem (Q1-2) to obtain the optimal transmit beamformer vectors $\mathbf{w}_{ij}$. Since problem (Q1-2) is convex w.r.t. each variable while keeping other variables fixed, problem (Q1-2) is much easier to solve than problem (P2). Specifically, problem (Q1-2) can be solved via the following alternating optimization procedure:

- For given $\mathbf{w}_{ij}$, $\forall\ i \in \mathcal{N}$ and $j \in \mathcal{L}$, the optimal receive

beamformer of problem (Q1-2) can be calculated by the well-known MMSE receiver:

$$y_i = \frac{\sum_{j\in\mathcal{L}} \mathbf{h}_{ij}^H \mathbf{w}_{ij}}{\sum_{k\in\mathcal{N}} \left(\sum_{j\in\mathcal{L}} \mathbf{h}_{ij}^H \mathbf{w}_{kj}\right)\left(\sum_{j\in\mathcal{L}} \mathbf{w}_{kj}^H \mathbf{h}_{ij}\right) + \sigma_i^2}. \tag{25}$$

- For fixed y_i and $\mathbf{w}_{ij}$, $\forall\ i \in \mathcal{N}$ and $j \in \mathcal{L}$, the optimal MSE weight x_i of problem (Q1-2) can be obtained by

$$x_i = \nabla\theta_i(e_i). \tag{26}$$

- For fixed x_i and y_i, $\forall\ i \in \mathcal{N}$, the optimal transmit beamformer vector $\mathbf{w}_{ij}$ can be obtained by solving the following quadratically constrained quadratic program (QCQP), which can be easily reformulated as a SOCP:

$$\begin{aligned} \min_{\mathbf{w}_{ij}} \quad & \sum_{i=1}^{N} x_i e_i + (P_f + \frac{1}{\eta}) \sum_{i=1}^{N}\sum_{j=1}^{L} \mathbf{w}_{ij}^H \mathbf{w}_{ij} \qquad (27)\\ \text{s.t.} \quad & \|\mathbf{r}_i\|_2 < \sqrt{1 + 1/(2^{\hat{c}_i} - 1)}\,\mathfrak{Re}[R_{ii}],\ \forall i \in \mathcal{N},\\ & \sum_{i=1}^{N} \mathbf{w}_{ij}^H \mathbf{w}_{ij} \le E_j,\ \forall j \in \mathcal{L}. \end{aligned}$$

where e_i is given by (24).

Therefore, we can solve problem (P2) with the iterative WMMSE method as elaborated in Algorithm 2, in which

$$O^{(p)} = \sum_{i=1}^{N} g_i(c_i^{(p)}) + (P_f + \frac{1}{\eta}) \sum_{i=1}^{N}\sum_{j=1}^{L} \left\|\mathbf{w}_{ij}^{(p)}\right\|_2^2,$$

and

$$c_i^{(p)} = B_i \log\left(1 + \frac{|\sum_{j\in\mathcal{L}} \mathbf{h}_{ij}^H \mathbf{w}_{ij}^{(p)}|^2}{\sigma_i^2 + \sum_{k\neq i}^{N} |\sum_{j\in\mathcal{L}} \mathbf{h}_{ij}^H \mathbf{w}_{kj}^{(p)}|^2}\right).$$

Algorithm 2 Iteratively WMMSE approach for the ESUM problem

1: Initialize: $\mathbf{w}_{ij}^{(0)}$ and $p = 1$.
2: **while** $\left|O^{(p)} - O^{(p-1)}\right| > \xi$ **do**
3: Given $\mathbf{w}_{ij}^{(p-1)}$, obtain receive beamformer $y_i^{(p)}$ by (25);
4: Fix $\mathbf{w}_{ij}^{(p-1)}$ and $y_i^{(p)}$, obtain the MSE weight $x_i^{(P)}$ from (26) and (24);
5: Given $x_i^{(p)}$, $y_i^{(p)}$ and $z_{ij}^{(p)}$, obtain the transmit beamformer $\mathbf{w}_{ij}^{(p)}$ by solving the convex optimization problem (27);
6: Update $c_i^{(p)}$;
7: Let $p = p + 1$.
8: **end while**
9: Output: $\mathbf{c}_o = [c_1^{(p)}, \cdots, c_N^{(p)}]^T$.

C. The RRH selection problem

After obtaining the achievable rates $\mathbf{c}^* = [c_1^*, \ldots, c_N^*]^T$ via solving problem (P2) by Algorithm 1 or Algorithm 2, we now turn to the RRH selection problem since the active RRH set $\mathcal{A}$ is relaxed to the full set $\mathcal{L}$ in the ESUM problem. The main focus of this subsection is to recover the active RRH set $\mathcal{A}$, based on the given optimal UE achievable rate vector $\mathbf{c}^*$ from Algorithm 1 or Algorithm 2.

Replacing μ_i and c_i in problem (P1) by the solutions μ_i^* and c_i^* obtained from Algorithm 1 respectively, we have the following RRH selection problem:

$$\begin{aligned} \min_{\mathbf{w}_{ij}} \quad & P_f \|\mathbf{m}\|_0 + \frac{1}{\eta} \sum_{i=1}^{N}\sum_{j=1}^{L} \mathbf{w}_{ij}^H \mathbf{w}_{ij} \qquad (28)\\ \text{s.t.} \quad & \sum_{i=1}^{N} \mathbf{w}_{ij}^H \mathbf{w}_{ij} \le E_j,\ \forall j \in \mathcal{L}.\\ & \|\mathbf{r}_i(\mathcal{L})\|_2 \le \sqrt{1 + 1/(2^{c_i^*/B_i} - 1)}\,\mathfrak{Re}[R_{ii}(\mathcal{L})],\\ & \qquad\qquad \forall i \in \mathcal{N}, \qquad (29) \end{aligned}$$

which is a MINLP.

We introduce an auxiliary binary variable $\beta_j \in \{0, 1\}, \forall j \in \mathcal{L}$, where $\beta_j = 1$ if and only if RRH j is active (the fiber link j is turned on). Then, problem (28) becomes

$$\begin{aligned} \min_{\beta_j, \mathbf{w}_{ij}} \quad & P_f \|\mathbf{m}\|_0 + \frac{1}{\eta} \sum_{i=1}^{N}\sum_{j=1}^{L} \mathbf{w}_{ij}^H \mathbf{w}_{ij} \qquad (30)\\ \text{s.t.} \quad & \sum_{i=1}^{N} \mathbf{w}_{ij}^H \mathbf{w}_{ij} \le \beta_j E_j,\ \forall j \in \mathcal{L}.\\ & \|\mathbf{r}_i(\mathcal{L})\|_2 \le \sqrt{1 + 1/(2^{c_i^*/B_i} - 1)}\,\mathfrak{Re}[R_{ii}(\mathcal{L})],\\ & \qquad\qquad \forall i \in \mathcal{N},\\ & \beta_j \in \{0, 1\},\ \forall j \in \mathcal{L},\\ & \sum_{j=1}^{L} \beta_j \ge 1, \qquad (31) \end{aligned}$$

where (31)) indicates that at least one RRH is turned on. Applying l_1-norm relaxation to $\|\mathbf{m}\|_0$ and the binary-to-continuous relaxation to the variable β_j, the relaxed RRH selection problem becomes

$$\begin{aligned} \text{(Q2)} \quad \min_{\beta_j, \mathbf{w}_{ij}} \quad & \sum_{j=1}^{L} \beta_j E_j\\ \text{s.t.} \quad & \|\mathbf{r}_i(\mathcal{L})\|_2 \le \sqrt{1 + 1/(2^{c_i^*/B_i} - 1)}\,\mathfrak{Re}[R_{ii}(\mathcal{L})],\\ & \qquad\qquad \forall i \in \mathcal{N},\\ & \sum_{i=1}^{N} \mathbf{w}_{ij}^H \mathbf{w}_{ij} \le \beta_j E_j,\\ & \sum_{j=1}^{L} \beta_j \ge 1,\\ & 0 \le \beta_j \le 1,\ \forall j \in \mathcal{L}, \end{aligned}$$

where the constraint $\sum_{j=1}^{L} \beta_j \ge 1$ can improve the accuracy of relaxation from problem (30) to (Q2), although it is redundant for problem (30). Problem (Q2) is a SOCP and can be solved easily by standard convex programming tool boxes [42]. Let the optimal solution of problem (Q2) be

$\{\tilde{\beta}_j, \tilde{w}_{ij} \mid i \in \mathcal{N},\ j \in \mathcal{L}\}$.

We can interpret $\tilde{\beta}_j$ to be the priority of RRH j being chosen to be active, where a RRH with relatively lower priority value should be turned off. However, [17] suggests that some incentive algorithms can help improve the RRH selection. We utilize the reweighted l_1-norm relaxation as the incentive strategy and propose the following *Shaping-and-Pruning (SP)* algorithm, which has two main steps:

1) **Shaping**. We use the reweighted l_1-norm relaxation [44] in (Q2) to "shape" the solutions into a sparse form. Specifically, we solve the reweighted problem

$$\begin{aligned} \min_{\beta_j, w_{ij}} \quad & \sum_{i=1}^{N}\sum_{j=1}^{L} \rho_j \beta_j E_j \\ \text{s.t.} \quad & \text{constraints in problem (Q2)}, \end{aligned} \tag{32}$$

where $\rho_j = 1/\left(\tilde{\beta}_j + \xi\right)$, ξ is adaptively chosen by $\xi = \max\left\{\min\left(\tilde{\beta}_1, \cdots, \tilde{\beta}_L\right), \phi\right\}$, and ϕ is a small positive value to ensure numerical stability [21]. We denote the optimal solution obtained from problem (32) as the *shaped priorities* $\{\hat{\beta}_j \mid j \in \mathcal{L}\}$.

2) **Pruning**. Sort the shaped priorities $\{\hat{\beta}_j \mid j \in \mathcal{L}\}$ in ascending order, so that $\hat{\beta}_{\pi_1} \leq \hat{\beta}_{\pi_2} \cdots \leq \hat{\beta}_{\pi_L}$, for some permutation $(\pi_1, \ldots, \pi_L)$ of the set $\mathcal{L}$. We define the Jth active RRH set to be $\mathcal{A}_J \triangleq \{\pi_{J+1}, \cdots, \pi_L\}$. Then, we apply the bisection search to find J^*, which is the largest index J such that $\beta_{\pi_1} = \cdots = \beta_{\pi_J} = 0$ and $\beta_{\pi_j} = \hat{\beta}_{\pi_j}$, for all $j \geq J+1$, form a feasible solution to (Q2). Finally, take the active set to be $\mathcal{A}^* = \mathcal{A}_{J^*}$.

After obtaining the active set $\mathcal{A}^*$, the corresponding beamforming weights $\mathbf{w}_{ij}^*$ can be found by solving the following SOCP:

$$\begin{aligned} \min_{\mathbf{w}_{ij}} \quad & \sum_{i=1}^{N}\sum_{j\in\mathcal{A}^*} \mathbf{w}_{ij}^H \mathbf{w}_{ij} \\ \text{s.t.} \quad & \sum_{i=1}^{N} \mathbf{w}_{ij}^H \mathbf{w}_{ij} \leq E_j,\ \forall j \in \mathcal{A}^*, \\ & \|\mathbf{r}_i(\mathcal{A}^*)\|_2 \leq \sqrt{1 + 1/(2^{c_i^*/B_i} - 1)}\,\mathfrak{Re}[R_{ii}(\mathcal{A}^*)], \\ & \qquad\qquad \forall i \in \mathcal{N}, \end{aligned} \tag{33}$$

where $\mathbf{w}_{ij}^* = 0,\ \forall i \in \mathcal{N}$, and $\forall j \notin \mathcal{A}^*$.

The proposed SP algorithm makes a trade-off between the conventional sorting-and-removing and sparsity-inducing algorithms [20], [22]. Specifically, the computational complexity is reduced by utilizing the bisection search, instead of the sequential and iterative search in sorting-and-removing algorithms. We summarize the Shaping-and-Pruning algorithm in Algorithm 3.

Remark 2: Note that, in the Shaping step in line 2 of Algorithm 3, we only need to solve problem (32) once, instead of iteratively updating the weights $\{\rho_j \mid j \in \mathcal{L}\}$, as is done in [20]. Suppose that the interior-point method is applied to solve the feasibility problem of (Q2), which is a SOCP, in each iteration. The time complexity to solve each SOCP is $\mathcal{O}((NLK)^{3.5})$,[2] where K is the number of antennas in each RRH [45]. The complexity of solving (32) and (33) are also both $\mathcal{O}((NLK)^{3.5})$. Therefore, the complexity of our SP algorithm is $\mathcal{O}((NLK)^{3.5} \log L)$.

Algorithm 3 Shaping-and-Pruning Algorithm

1: Initialization. Solve problem (Q2) to obtain $\{\tilde{\beta}_j \mid j \in \mathcal{L}\}$. Let $J_{\min} = 0$ and $J_{\max} = L$.
2: Shaping: Solve problem (32), where ρ_j, $j = 1, \ldots, L$, are defined by $\{\tilde{\beta}_j \mid j \in \mathcal{L}\}$, to obtain the shaped priorities $\{\hat{\beta}_j \mid j \in \mathcal{L}\}$.
3: Pruning: Sort the shaped priorities $\{\hat{\beta}_j \mid j \in \mathcal{L}\}$ in ascending order, to obtain $\hat{\beta}_{\pi_1} \leq \hat{\beta}_{\pi_2} \cdots \leq \hat{\beta}_{\pi_L}$.
4: **while** $J_{\max} - J_{\min} \geq 2$ **do**
5: $J = \lfloor (J_{\max} + J_{\min})/2 \rfloor$;
6: Check the feasibility of problem (Q2) if $\mathcal{A} = \mathcal{A}_J$;
7: **if** feasible **then**
8: $J_{\min} = J$;
9: **else**
10: $J_{\max} = J$;
11: **end if**
12: **end while**
13: Output $J^* = \lfloor (J_{\max} + J_{\min})/2 \rfloor$, $\mathcal{A}^* = \mathcal{A}_{J^*} = \{\pi_{J^*+1}, \cdots, \pi_L\}$ and its corresponding beamforming weights $\mathbf{w}_{ij}^*$ by solving (33).

In summary, the proposed solution for problem (P1) is obtained as follows: the optimal VM computation capacities and achievable rates $\{(\mu_i^*, c_i^*) \mid i \in \mathcal{N}\}$ are obtained from Algorithm 1 or Algorithm 2, which are then used to determine the optimal active RRH set $\mathcal{A}^*$ and its corresponding beamforming weights w_{ij}^* from Algorithm 3. For ease of reference, we call the whole procedure that solve problem (P1) by Algorithm 1 and Algorithm 3 in tandem as Cross-Layer Shaping-and-Pruning algorithm 1 (CLSP1) and the whole procedure that solve problem (P1) by Algorithm 2 and Algorithm 3 in tandem as Cross-Layer Shaping-and-Pruning algorithm 2 (CLSP2).

V. Simulation Results

In this section, we present simulation results to verify the performance of the proposed CLSP1 and CLSP2 algorithms, and compare them to several existing algorithms in the literature.

A. Simulation setup

We consider a heterogeneous C-RAN system of 7 RRHs, where RRH 1 to 6 are located on a circle centered at a macro RRH, with radius 0.5 km. The RRHs 1 to 6 are placed at equal distances apart, as shown in Figure 4. The maximum transmitting power from RRH 1 to RRH 6 is $E_1 = E_2 = \cdots = E_6 = E$, and the maximum transmitting power for the macro RRH is E_m. The wireless transmission bandwidth is 10 MHz. We adopt the path loss model used by the 3GPP specification for Evolved Universal Terrestrial Radio Access

[2] We say that $q(n) = \mathcal{O}(g(n))$ if $\limsup_{n\to\infty} |q(n)/g(n)| < \infty$.

TABLE I
SIMULATION PARAMETERS

Parameter	Value	Parameter	Value
L	7	K	2
σ^2	-83.98 dBm	η	0.2
E	1 W	E_m	10 W
ϑ	5 dB	s	10 dB
P_f	5 W		

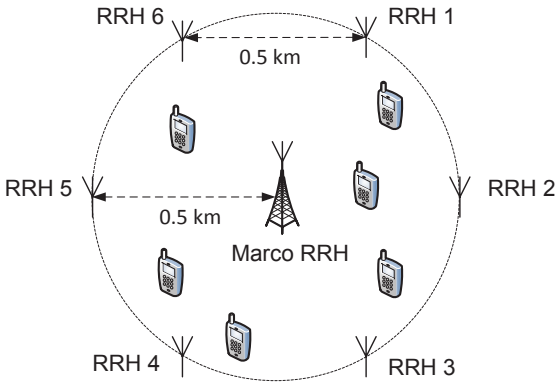


Fig. 4. Simulation setup in a heterogeneous C-RAN.

in [46], where the received power at a UE d km from a RRH is given by

$$p \text{ (dB)} = 128.1 + 37.6 \log_{10} d.$$

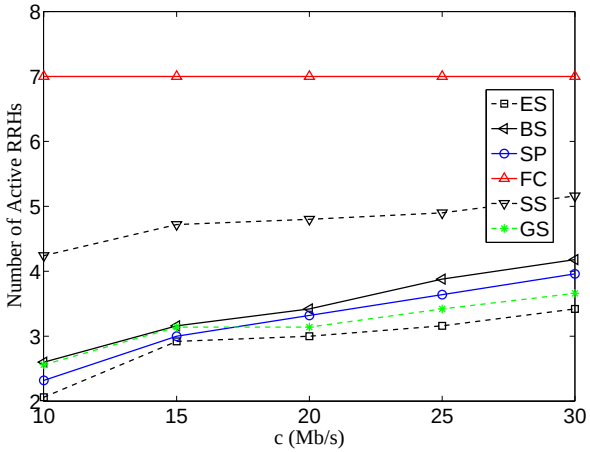
The transmit antenna gain at each RRH is ϑ. The lognormal shadowing parameter is s. In our simulations, we consider homogeneous UEs with $\sigma_1 = \sigma_2 = \cdots = \sigma_N = \sigma$, and $\tau_1 = \tau_2 = \cdots = \tau_N = \tau$.

For the power consumption function $\varphi_i(\mu_i)$, we adopt the formula $\varphi_i(\mu_i) = k_i \mu_i^3$, where $k_i > 0$ is a constant. This power consumption formula was proposed by [35] and adopted by [36], [37]. We summarize our simulation parameters in Table I [20], [29].

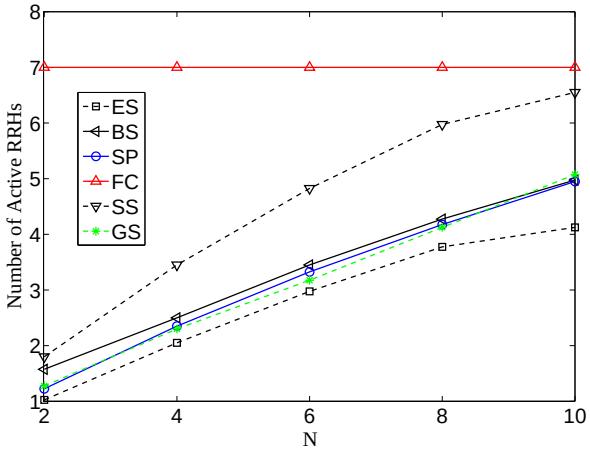
B. The effect of shaping

In this subsection, we show the performance of the proposed SP algorithm for the RRH selection problem, compared with the following benchmark algorithms:

- *Exhaustive Search (ES) Algorithm.* This algorithm solves the RRH selection problem (28) using an exhaustive search over all possible RRH selections to obtain the optimal solution for problem (28). It has a high complexity of $\mathcal{O}((NLK)^{3.5} 2^L)$, which makes the algorithm intractable when L becomes large. This is used as a benchmark to compare other algorithms against.
- *Bisection Search (BS) Algorithm.* This algorithm, which was proposed in [47], skips the shaping step in the SP algorithm, and uses $\tilde{\beta}_j$ in place of $\hat{\beta}_j$ for all $j \in \mathcal{L}$ in the pruning step of the SP algorithm. We use this algorithm as a benchmark to show the effect of the shaping step in the SP algorithm. The complexity of the BS algorithm is $\mathcal{O}((NLK)^{3.5} \log L)$.



(a) Different required achievable rates.



(b) Different number of UEs.

Fig. 5. Number of active RRHs using different RRH selection algorithms.

- *Full Cooperation (FC) Algorithm.* This algorithm assumes all the RRHs are chosen to be active, i.e., $\mathcal{A}^* = \mathcal{L}$. The complexity of the FC algorithm is $\mathcal{O}((NLK)^{3.5})$.
- *Successive Selection (SS) Algorithm.* This algorithm was proposed in [29], and lets all RRHs to be active in the initial iteration, with a RRH having the least power consumption removed at each subsequent iteration. The iterations are performed until the problem (28) becomes infeasible. The complexity of the SS algorithm is $\mathcal{O}((NLK)^{3.5} L)$.
- *Greedy Selection (GS) Algorithm.* This algorithm was proposed in [22]. It considers all RRHs to be active in the initial iteration, and then removes the RRH that reduces the system power consumption by the largest amount at each iteration until the problem (28) becomes infeasible. Simulation results in [22] suggest that this algorithm produces a near-optimal solution, compared with the global solution obtained by solving a MINLP. The complexity of the GS algorithm is $\mathcal{O}((NLK)^{3.5} L^2)$.

To compare the performance of all the RRH selection algorithms, we suppose the optimal achievable rate c_i^* for each

UE i is identical in problem (28), i.e., $c_1^* = \cdots = c_N^* = c$. In Figure 5(a), we show the mean number of active RRHs versus each UE's optimal achievable rate c under different RRH selection algorithms when the number of UEs $N = 6$. We see that the SP algorithm outperforms the FC, SS and BS algorithms over all mean arrival rates. When $c \leq 17$ Mb/s, the SP algorithm has comparable or even better sparsity performance than the GS algorithm. When $c \geq 17$ Mb/s, compared to the SP algorithm, the GS algorithm produces a solution with about 5% less active RRHs, but at the expense of $L^2/\log L = 17.5$ times computational complexity overhead.

Next, we let $c = 20$ Mb/s, and show the mean number of active RRHs versus the number of UEs N under different RRH selection algorithms in Figure 5(b). We see from Figure 5(b) that the SP algorithm has similar sparsity performance as the GS algorithm, and outperforms the FC, SS and BS algorithms. From both Figures 5(a) and 5(b), we can see that although we incur an overhead to perform the shaping step in the SP algorithm, its solution sparsity is improved by 5% - 10%.

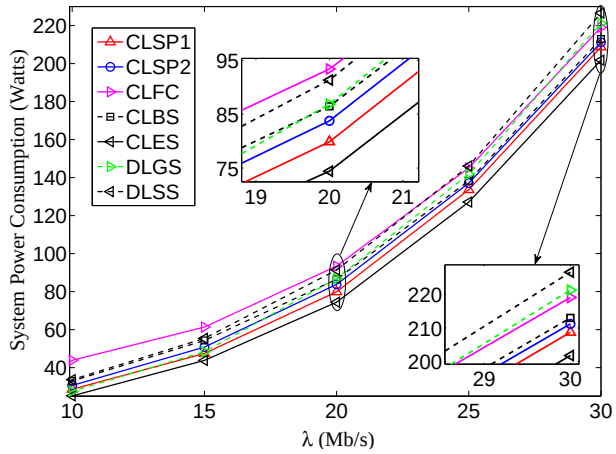
C. The importance of cross-layer design

In this subsection, we present simulation results to verify the performance gain using a cross-layer design in which the both the BBU pool power consumption and the RRH power consumption are jointly optimized. Most of the previous work in C-RAN optimizes the power consumption for the wireless transmission layer and BBU pool independently, for instance, [29] and [22]. We call this class of algorithms the decoupled-layer (DL) algorithms. We assume that, for the DL algorithms, the delay in the BBU processing queue a_i and the delay in the RRH transmitting queue b_i satisfy $a_i \leq \tau_i/2$ and $b_i \leq \tau_i/2$ respectively. We formulate optimization programs, similar to problem (P0), for finding optimal UE achievable rates, VM computation capacities, beamformer vectors, and active RRH set separately. The RRH selection problem can then be solved using either the SS or GS algorithms. We call these the DLSS and DLGS method respectively.

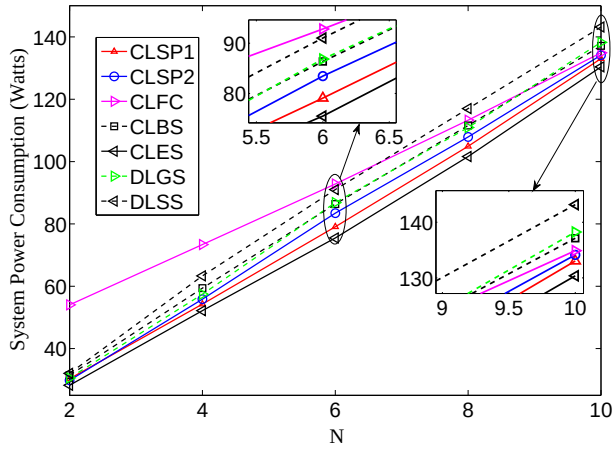
In this paper, we have provided a general framework that allows us to perform cross-layer (CL) optimization of the overall system power consumption. Our two-step approach (cf. Figure 3) allows us to first solve a QWSRM using Algorithm 1 and then a RRH selection problem. For the RRH selection problem, we can again adopt the ES, BS, and FC algorithms. We call these the CLES, CLBS, and CLFC methods respectively[3]. We let UEs' mean arrival rates to be identical, i.e., $\lambda_1 = \cdots = \lambda_N = \lambda$.

In Figure 6, we show the relationship between the UEs' mean arrival rate and the system power consumption for $N = 6$ in Figure 6(a). We observe that, firstly, CL algorithms outperform the DL algorithms, especially in high traffic rate regime. Secondly, CLSP1 outperforms CLSP2. Finally, as the incoming traffic rate increases, the performance gap between CLFC and CLSP1 (CLSP2) becomes smaller since CLSP1 (CLSP2) needs more active RRHs to support the higher rate demand. The performance of system power consumption versus the number of UEs is depicted in Figure 6(b) when $\lambda = 20$ Mb/s. We see again that CL algorithms are better than DL algorithms. CLSP1 and CLSP2 also outperforms CLBS, which again shows the importance of the shaping step in the SP algorithm.

[3]Since SS and GS have high complexity and have performance lower bounded by ES, we do not include these methods in tandem with Algorithm 1 as our benchmarks for comparison.



(a) Different arrival rates.



(b) Different number of UEs.

Fig. 6. System power consumption under different algorithms.

VI. Conclusion

In this paper, we have investigated the problem of minimizing the overall system power consumption (including the power consumption in the BBU pool, the fiber links and the RRHs) in a C-RAN, such that the cross-layer QoS and per-RRH power constraints are satisfied. We formulated a MINLP and then relax it to an ESUM problem, which gives the optimal achievable rate for each UE. Based on the optimal achievable rate, we proposed an efficient SP algorithm, with lower computational complexity than several state-of-the-art RRH selection methods, to recover a sparse solution for the RRH selection problem. Simulation results suggest that our proposed SP algorithm outperforms various other methods,

and the proposed cross-layer algorithm is more energy efficient than existing decoupled-layer methods.

C-RAN provides a centralized BBU pool, instead of the distributed BSs, to improve resource utilization, and enable the use of centralized processing like joint beamforming. However, there are two main side effects. The first is the large channel state information (CSI) overhead, and the second is the high amount of data transfer in the fronthaul, whose capacity is limited in practice. As the number of RRHs and UEs in C-RANs becomes large, our proposed algorithms may be restricted by the large CSI overhead and limited fronthaul capacity, and may become unsuitable for real-time implementation. Therefore, in future work, it would be of interest to incorporate CSI overhead reduction techniques based on historical traffic arrival rates, and statistical properties of the channel states, in order to perform approximate real-time cross-layer optimization in limited fronthaul capacity systems.

APPENDIX A
PROOF OF THEOREM 1

Suppose that there exists some $i \in \mathcal{N}$ such that $\tilde{c}_i < \bar{c}_i$, and $c_i^* \in (\tilde{c}_i, \bar{c}_i]$. Since $f(\mathbf{c})$ is convex and finite, there exists $\hat{c}_i \in (\tilde{c}_i, c_i^*)$ such that $f(\hat{\mathbf{c}}) < f(\mathbf{c}^*)$, where $\hat{\mathbf{c}}$ is $\mathbf{c}^*$ with the i-th element c_i^* replaced by $\hat{c}_i$. In addition, we have

$$\begin{aligned}\|\mathbf{r}_i^*(\mathcal{A})\|_2 &\le \sqrt{1+1/(2^{c_i^*/B_i}-1)}\,\mathfrak{Re}[R_{ii}^*(\mathcal{A})] \\ &< \sqrt{1+1/(2^{\hat{c}_i/B_i}-1)}\,\mathfrak{Re}[R_{ii}^*(\mathcal{A})],\end{aligned}$$

which implies that $\hat{\mathbf{c}}$ is a feasible rate vector for the QWSRM (14). But $f(\hat{\mathbf{c}}) < f(\mathbf{c}^*)$, which contradicts the assumption that $\mathbf{c}^*$ is optimal for the QWSRM. The theorem is now proved.

APPENDIX B
PROOF OF PROPOSITION 1

If we fix the variables $\{\mu_i, c_i\}$ in problem (P1), then problem (P1) is reduced to

$$\begin{aligned}\min_{\mathbf{w}_{ij}} \quad & \sum_{i=1}^{N}\sum_{j=1}^{L}\mathbf{w}_{ij}^H\mathbf{w}_{ij} \qquad (34)\\ \text{s.t.} \quad & \|\mathbf{r}_i(\mathcal{L})\|_2 \le \sqrt{1+\frac{1}{2^{c_i/B_i}-1}}\,\mathfrak{Re}[R_{ii}(\mathcal{L})], \\ & \qquad\qquad \forall i \in \mathcal{N}, \\ & \sum_{i=1}^{N}\mathbf{w}_{ij}^H\mathbf{w}_{ij} \le E_j,\ \forall j \in \mathcal{L},\end{aligned}$$

which is a SOCP. Then, we can observe that, if we slightly increase the value of constant c_i, the feasible region of problem (34) is shrunk accordingly. That means the optimal value of problem (34) is nondecreasing w.r.t. c_i.

On the other hand, from Assumption 1, $\varphi_i(\mu_i)$ is increasing w.r.t. μ_i. Therefore, the optimal $\{\mu_i, c_i\}$ of problem (P1) must achieve equality in the system delay constraint (5) since the left hand side of (5) is monotonically decreasing w.r.t. μ_i and c_i respectively. The proposition is now proved.

APPENDIX C
PROOF OF PROPOSITION 2

On the one hand, from Assumption 1, $\varphi_i(\cdot)$ is convex and increasing; on the other hand, $\lambda_i + \frac{1}{\tau_i} + \frac{1}{\tau_i^2(c_i-\lambda_i)-\tau_i}$ is convex w.r.t. $c_i > \lambda_i + 1/\tau_i$. Then it can be shown that for each $i \in \mathcal{N}$, if $c_i > \lambda_i + 1/\tau_i$, $f_i(c_i)$ is convex, based on the composition rules, which preserve convexity [48].

Therefore, $\sum_{i=1}^{N} f_i(c_i)$ is a convex function over $\hat{\mathcal{Q}}_{\text{init}}$, and it also satisfies the three properties of the objective function $f(\mathbf{c})$ in a QWSRM problem (14) in Section III. Therefore, problem (Q1-1) is a QWSRM problem. In addition, $\bar{c}_i$ is an upper bound of c_i derived from (22), since for any $i \in \mathcal{N}, j \in \mathcal{L}$, we have $\|\mathbf{w}_{ij}\|_2^2 \le E_j$. The proposition is now proved.

REFERENCES

[1] China Mobile Research Institute, "C-RAN: The road towards green RAN," China Mobile Research Institute, White Paper, V2.5, Oct. 2011.

[2] C.-L. I, C. Rowell, S. Han, Z. Xu, G. Li, and Z. Pan, "Toward green and soft: a 5G perspective," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 66–73, Feb. 2014.

[3] W. H. Chin, Z. Fan, and R. Haines, "Emerging technologies and research challenges for 5G wireless networks," *IEEE Wireless Commun. Mag.*, vol. 21, no. 2, pp. 106–112, Apr. 2014.

[4] P. Rost, C. J. Bernardos, A. D. Domenico, M. D. Girolamo, M. Lalam, A. Maeder, D. Sabella, and D. W$\ddot{\text{u}}$bben, "Cloud technologies for flexible 5G radio access networks," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 68–76, May 2014.

[5] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, "What will 5G be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[6] B. Sourjya, C. S. Preeth, J. M. Kashyap, K. Gautam, M. Anand, P. Paul, S. Vikram, and W. Thomas, "CloudIQ: A framework for processing base stations in a data center," in *Proc. ACM MobiCom*, Istanbul, Turkey, Aug. 2012, pp. 125–136.

[7] K. Sundaresan, M. Y. Arslan, S. Singh, S. Rangarajan, and S. V. Krishnamurthy, "FluidNet: A flexible cloud-based radio access network for small cells," in *Proc. ACM MobiCom*, Miami, FL, Sep. 2013, pp. 99–110.

[8] Y. Mao, L. Yong, J. Depeng, S. Li, M. Shaowu, and Z. Lieguang, "OpenRAN: A software-defined RAN architecture via virtualization," in *Proc. ACM SIGCOMM*, Hong Kong, China, Aug. 2013, pp. 549–550.

[9] C. Liu, K. Sundaresan, M. Jiang, S. Rangarajan, and G.-K. Chang, "The case for re-configurable backhaul in cloud-RAN based small cell networks," in *Proc. IEEE INFOCOM*, Turin, Italy, Apr. 2013, pp. 1124–1132.

[10] D. Sabella, P. Rost, Y. Sheng, E. Pateromichelakis, U. Salim, P. Guitton-Ouhamou, M. D. Girolamo, and G. Giuliani, "RAN as a service: Challenges of designing a flexible RAN architecture in a cloud-based heterogeneous mobile network," in *Proc. Future Network and Mobile Summit*, Lisbon, Portugal, Jul. 2013, pp. 1–8.

[11] D. Feng, C. Jiang, G. Lim, L. J. J. Cimini, G. Feng, and G. Y. Li, "A survey of energy-efficient wireless communications," *IEEE Commun. Surveys & Tutorials*, vol. 15, no. 1, pp. 167–178, 2013.

[12] Y. Zeng, E. Gunawan, Y. L. Guan, and J. Liu, "Joint base station selection and linear precoding for cellular networks with multi-cell processing," in *Proc. IEEE TENCON*, Fukuoka, Japan, Nov. 2010, pp. 1976–1981.

[13] Q. Du and X. Zhang, "QoS-aware base-station selections for distributed MIMO links in broadband wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 6, pp. 1123–1138, Jun. 2011.

[14] M. Hong and Z.-Q. Luo, "Distributed linear precoder optimization and base station selection for an uplink heterogeneous network," *IEEE Trans. Signal Process.*, vol. 61, no. 12, pp. 3214–3228, Jun. 2013.

[15] D. Amzallag, R. Bar-Yehuda, D. Raz, and G. Scalosub, "Cell selection in 4G cellular networks," *IEEE Trans. Mobile Comput.*, vol. 12, no. 7, pp. 1443–1455, Jul. 2013.

[16] C.-P. Chien, K.-M. Yang, and H.-Y. Hsieh, "Selection of transmission points for delay minimization in LTE-A heterogeneous networks with low-power RRHs," in *Proc. IEEE WCNC*, Shanghai, China, Apr. 2013, pp. 783–788.

[17] Y. Cheng, M. Pesavento, and A. Philipp, "Joint network optimization and downlink beamforming for CoMP transmissions using mixed integer conic programming," *IEEE Trans. Signal Process.*, vol. 61, no. 16, pp. 3972–3987, Aug. 2013.

[18] A. Liu and V. Lau, "Joint power and antenna selection optimization for energy-efficient large distributed MIMO networks," in *Proc. IEEE ICCS*, Singapore, Nov. 2012, pp. 230–234.

[19] S.-H. Park, O. Simeone, O. Sahin, and S. Shamai, "Robust and efficient distributed compression for cloud radio access networks," *IEEE Trans. Veh. Technol.*, vol. 62, no. 2, pp. 692–703, Feb. 2013.

[20] J. Zhao, T. Q. S. Quek, and Z. Lei, "Coordinated multipoint transmission with limited backhaul data transfer," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2762–2775, Jun. 2013.

[21] B. Dai and W. Yu, "Sparse beamforming for limited-backhaul network mimo system via reweighted power minimization," in *Proc. IEEE GLOBECOM*, Atlanta, GA, Dec. 2013, pp. 1962–1967.

[22] Y. Shi, J. Zhang, and K. B. Letaief, "Group sparse beamforming for green cloud-RAN," *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, pp. 2809–2823, May 2014.

[23] S. Luo, R. Zhang, and T. J. Lim, "Downlink and uplink energy minimization through user association and beamforming in C-RAN," *IEEE Trans. Wireless Commun.*, vol. 14, no. 1, pp. 494–508, Jan. 2015.

[24] J. Bartelt, G. Fettweis, D. Wubben, M. Boldi, and B. Melis, "Heterogeneous backhaul for cloud-based mobile networks," in *Proc. IEEE VTC*, Las Vegas, NV, Sep. 2013, pp. 1–5.

[25] M. C. Valenti, S. Talarico, and P. Rost, "The role of computational outage in dense cloud-based centralized radio access networks," in *Proc. IEEE GLOBECOM*, Austin, TX, Dec. 2014, pp. 1489–1495.

[26] 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, no. 22, pp. 5646–5658, Nov 2013.

[27] S. Joshi, P. Weeraddana, M. Codreanu, and M. Latva-aho, "Weighted sum-rate maximization for MISO downlink cellular networks via branch and bound," *IEEE Trans. Signal Process.*, vol. 60, no. 4, pp. 2090–2095, Apr. 2012.

[28] Q. Shi, M. Razaviyayn, Z.-Q. 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.

[29] Y. Shi, J. Zhang, and K. B. Letaief, "Group sparse beamforming for green cloud radio access networks," in *Proc. IEEE GLOBECOM*, Atlanta, GA, 2013, pp. 4635–4640.

[30] D. Wu and R. Negi, "Effective capacity: a wireless link model for support of quality of service," *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, pp. 630–643, Jul. 2003.

[31] P. J. Burke, "The output of a queuing system," *Operations Research*, vol. 4, no. 6, pp. 699–704, Dec. 1956.

[32] E. Reich, "Waiting times when queues are in tandem," *The Annals of Mathematical Statistics*, vol. 28, no. 3, pp. 768–773, Sep. 1957.

[33] D. Bertsekas and R. Gallager, *Data Networks*, 2nd ed. New Jersey, U.S.: Prentice Hall, 1992.

[34] D. W. H. Cai, T. Q. S. Quek, C. W. Tan, and S. H. Low, "Max-min SINR coordinated multipoint downlink transmission — duality and algorithms," *IEEE Trans. Signal Process.*, vol. 60, no. 10, pp. 5384–5395, Oct. 2012.

[35] S. Kaxiras and M. Martonosi, *Computer Architecture Techniques for Power-Efficiency*, 1st ed. Morgan and Claypool Publishers, 2008.

[36] L. Chen, N. Li, and S. H. Low, "On the interaction between load balancing and speed scaling," in *Proc. ITA Workshop*, San Diego, CA, Feb. 2011.

[37] J. Tang, W. P. Tay, and Y. Wen, "Dynamic request redirection and elastic service scaling in cloud-centric media networks," *IEEE Trans. Multimedia*, vol. 16, no. 5, pp. 1434–1445, Aug. 2014.

[38] 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, Sep. 2013.

[39] A. Wiesel, Y. C. Eldar, and S. S. (Shitz), "Linear precoding via conic optimization for fixed MIMO receivers," *IEEE Trans. Signal Process.*, vol. 54, no. 1, pp. 161–176, Jan. 2006.

[40] X. Lin, N. B. Shroff, and R. Srikant, "A tutorial on cross-layer optimization in wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, pp. 1452–1463, Aug. 2006.

[41] X. Zheng, F. Chen, Y. Xia, and Y. Fang, "A class of cross-layer optimization algorithms for performance and complexity trade-offs in wireless networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 20, no. 10, pp. 1393–1407, Oct. 2009.

[42] CVX Research, Inc., "CVX: Matlab software for disciplined convex programming, version 2.0," http://cvxr.com/cvx, Aug. 2012.

[43] S. Boyd and J. Mattingley, "Branch-and-bound methods," Mar. 2007. [Online]. Available: http://see.stanford.edu/materials/lsocoee364b/17-bb_notes.pdf

[44] E. J. Candès, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted l_1 minimization," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5, pp. 877–905, 2008.

[45] Y.-F. Liu, Y.-H. Dai, and Z.-Q. Luo, "Coordinated beamforming for MISO interference channel: Complexity analysis and efficient algorithms," *IEEE Trans. Signal Process.*, vol. 59, no. 3, pp. 1142–1157, Mar. 2011.

[46] 3GPP, "LTE; Evolved universal terrestrial radio access (E-UTRA); Radio frequency (RF) requirements for LTE Pico Node B (release 9)," 3rd Generation Partnership Project (3GPP), TS 36.931, May 2011, v9.0.0.

[47] J. Tang, W. P. Tay, and T. Q. S. Quek, "Cross-layer resource allocation in cloud radio access network," in *Proc. IEEE GlobalSIP*, Atlanta, GA, Dec. 2014, pp. 313–317.

[48] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge University Press, 2004.