Delay Aware Resource Allocation with Radio Remote Head Cooperation in user-centric C-RAN

Nazanin Moosavi, Mahnaz Sinaie, Paeiz Azmi, Jyrki Huusko

***Abstract*—High spectral efficiency and low latency are required to provide ubiquitous communication for the emerging applications in 5G wireless communication networks. In this letter, we propose a novel framework that considers these requirements simultaneously by integrating the notion of effective capacity (EC) into orthogonal frequency division multiple access (OFDMA) cloud-radio access networks (C-RAN) where the users select the distributed radio remote heads (RRHs) based on their specific delay requirements to transmit over different subcarriers cooperatively. Consequently, an optimization problem is defined to maximize the EC under the average peak power constraint and the delay requirements. The problem is combinatorial and non-convex and an algorithm based on the duality and alternating optimization algorithms is proposed, which is efficiently computed with good accuracy. Simulation and analytical results demonstrate that the proposed solution has a near-optimal performance and there is a trade-off between delay and spectral efficiency. Moreover, the cooperation between RRHs can considerably improve the system throughput.**

***Index Terms*- Effective Capacity, delay, user-centric C-RAN, CoMP**

I. Introduction

In recent years, there are growing demands for ubiquitous high-speed wireless access and the explosive proliferation of smartphones. The increasing demands make it more challenging for the network providers to manage and operate wireless networks to provide the required delay efficiently [1]. As a promising new technology and architecture for the future 5G standard, 5G cloud radio access network (C-RAN) has drawn significant attention in both industry and academia. It provides high flexibility, tremendous capacity, wide-coverage, and cost-effective operation mainly by incorporating powerful cloud computing and virtualization techniques [2]. C-RAN enables a centralized processing architecture in the baseband unit that provides joint signal processing for a cluster of cost-efficient low power base stations, called radio remote heads (RRHs), and their served users. It facilitates the implementation of coordinated multipoint transmission (CoMP). There exist two RRH clustering models, cell-centric that each RRH forms a cluster to serve the selected users, and user-centric that allows each user to associate with a set of RRHs. The most important benefit of employing the user-centric comparing to cell-centric is that the network becomes capable of providing higher throughput. Because in cell-centric, the users at the cluster edges suffer from considerable inter-cluster interference. However, by applying the user-centric, each user is served by an individually selected subset of neighboring RRHs. Hence, there exists no explicit cell edge, and the resources are allocated based on user groups of different service types, which results in better performances compared to the cell-centric models. C-RAN architecture is one of the suitable architectures to implement user-centric. Because the baseband unit is responsible for supporting the entire network and it controls a large number of radio units. Hence, it is capable of forming globally optimal user-centric clusters. Consequently, we consider the user-centric RRH clustering in our system model [3]–[6].

Although C-RAN has many benefits, it faces some challenges in strict delay requirements and fronthaul link capacity. Consequently, the 3GPP introduced a new approach called function splitting [7], [8] that the baseband is divided into a distributed unit (DU) and a centralized unit (CU). Based on the application with various rate requirements, different functions are executed in DU and CU. Besides this technique, efficient resource allocation algorithms must be designed to meet these challenges. Resource allocation for C-RAN has been extensively investigated in the following specific areas: RRH clustering and beamforming, physical resource block allocation, and deterministic delay optimization [9], [10]. There are some papers that considered delay in C-RAN [11]–[13]. The mentioned articles considered the delay analysis in the C-RAN network by applying different deterministic, stochastic and, statistical methods such as Lyapunov, Markov decision process and, effective capacity. However, due to variations of a wireless channel, guaranteeing the deterministic delay is challenging. Hence, the statistical delay, where the end-to-end delay is bounded with a certain violation probability, is suitable for real-time traffics over wireless channels. Meanwhile, the statistical delay does not require the dynamic status of the network queue for optimal resource allocations. Hence, effective capacity is introduced which is an alternative performance metric that takes into account the statistical delay of communication. Hence, it appears more suitable when both the system data-rate and communication statistical delay are the goal of optimization [14], [15]. In [16], an analytical framework is proposed to characterize the distribution of queuing delays at an aggregation gateway in the uplink. A delay exponent approach was used to satisfy the delay constraints of different service classes in [17]. Nevertheless, they ignored maximum peak power constraint, limited fronthaul capacity,

Nazanin Moosavi and Paeiz Azmi are with the Department of Electrical Engineering, Tarbiat Modares University, Tehran, Iran, Email:{n_moosavi,pazmi}@modares.ac.ir. Mahnaz Sinaie and JyrkiHuusko are with the VTT Technical Research Centre of Finland, Espoo, Finland, Email: {Mahnaz.sinaie,Jyrki.Huusko}@vtt.fi. The work of Mahnaz Sinaie was supported by Iran National Science Foundation (INSF) under grant 96007791 and It is also supported by ERCIM program. The work of Paeiz Azmi was supported by Iran National Science Foundation (INSF) under grant 91059145.

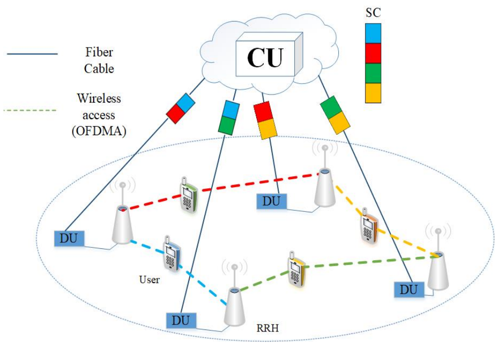

Figure 1: System Model

RRH cooperation, and multiuser interference management. In [18], the authors considered the effective capacity theory for the statistical delay provisioning in a simple two RRH scenario without considering the CoMP scheme and the limited fronthaul capacity. One of the important constraints in the context of delay analysis in the C-RAN network, which is not investigated in any of these papers, is CoMP selection based on users' delay requirements. Consequently, we consider the notion of effective capacity in OFDMA-based C-RAN networks for joint resource allocation and CoMP selection to guarantee the users' delay requirements.

In this letter, we propose a cross-layer radio resource allocation based on effective capacity in OFDMA-based C-RAN by considering the statistical delay requirements. We study a user-centric approach where users select the groups of RRHs based on their specific delay requirements to cooperatively transmit over different subcarriers. The optimization problem is non-convex due to the binary variables. An efficient solution is proposed based on duality and greedy algorithms. The various simulation setups demonstrate that the proposed solution with lower complexity has a very near-optimal performance. Furthermore, the results reveal that the power allocation strategy depends on not only channel gains but also the users' delay requirements. The simulation results demonstrate that there is a trade-off between delay and spectral efficiency and, the CoMP selection significantly increases the system throughput.

II. System Model and Problem Formulation

We consider a two-tier downlink OFDMA-based 5G C-RAN network which is illustrated in Fig.1. Let $\mathcal{M}=\{1,2,...,M\}$ denote the set of RRHs and $\mathcal{K}$=$\{1,2,...,K\}$ denote the set of users in the network. One CU-pool with fiber cable is connected to M DUs and each DU is connected to only one RRH. Moreover, RRH cooperation is also considered. Channel bandwidth is B MHz which is equally divided into N orthogonal subcarriers (SCs) and $\mathcal{N}=\{1,2,...,N\}$ is the set of SCs. Let $\phi_{k,n}$ denote the subcarrier assignment strategy, which is defined as follows:

$$\text{C1:}\quad \phi_{k,n}=\begin{cases} 1 & \text{If SC } n \text{ is assigned to the user } k,\\ 0 & \text{otherwise.}\end{cases} \tag{1}$$

Each SC $n\in\mathcal{N}$ is assigned to no more than one user, then:

$$\text{C2:}\quad \sum_{k=1}^{K}\phi_{k,n}=1, \forall n\in\mathcal{N}. \tag{2}$$

Hence, the set of SCs assigned to user k is given by $\mathcal{N}_k=\{n|\phi_{k,n}=1\}$, where $\mathcal{N}=\{\mathcal{N}_1\cup....\cup\mathcal{N}_K\}$ and $\mathcal{N}_i\cap\mathcal{N}_k=\phi,\ \forall i,k\in\mathcal{N}$. Based on CoMP scheme, each particular user selects a subset of RRHs on each SC. As a result, each RRH m transmits the received data from the DU only on the subset of SCs that are assigned to the users. Let $\alpha_{m,n}$ denotes the CoMP selection on SC n as follows:

$$\text{C3:}\quad \alpha_{m,n}=\begin{cases} 1 & \text{If RRH}\, m \text{ transmits data on SC}\, n,\\ 0 & \text{otherwise.}\end{cases} \tag{3}$$

Then, we define the subset of RRHs that transmit on SC $n\forall n\in\mathcal{N}$ as $\mathcal{A}_n=\{m|\alpha_{m,n}=1\}$. Thus, the RRHs in $\mathcal{A}_n$ cooperatively send the data to the user k over SC n.

For delay-sensitive services, delay provisioning is an important and challenging topic for beyond 5G networks. However, guaranteeing the deterministic delays for the mobile services is difficult and impractical due to the time-varying nature of the wireless channels. The concept of effective capacity (EC) is proposed in [14] to meet this challenge. It is defined as the maximum constant arrival rate over time-varying channels under a statistical delay requirement, which is specified by the delay exponent θ. The EC is formulated as $E_c=-\frac{1}{\theta_k}\ln\left(\mathbb{E}\{e^{-\theta R}\}\right)$, where $\mathbb{E}$ is the statistical expectation with respect to the service rate $R=\log_2(1+\text{SNR})$ in which SNR is modeled as a random variable due to the fading channel. Hence, the effective capacity is a generalization of the Shannon capacity, which accounts jointly the communication delay and reliability. The service rate of user k on $\mathcal{N}_k$ is R_k=$\frac{B}{N}\sum_{n\in\mathcal{N}_k}\log_2(1+\gamma_{k,n})$, where $\gamma_{k,n}$=$\frac{\sum_{m=1}^{M}\alpha_{m,n}p_{k,m,n}|g_{k,m,n}|^2}{\sigma^2}$. Here $g_{k,m,n}$, $p_{k,m,n}$, and σ^2 denote the complex wireless access channel coefficient, transmit power to the user k from RRH m on SC n and the receiver noise power variance respectively. Plugging service rate into effective capacity formulation yields the effective capacity of the user k, namely

$$\begin{aligned} E_c^k &= \frac{-1}{\theta_k}\log\mathbb{E}_g\Big\{e^{-\theta_k\sum_{n\in\mathcal{N}_k}\frac{B}{N}\log_2(1+\gamma_{k,n})}\Big\}\\ &=\frac{-1}{\theta_k}\log\mathbb{E}_g\left\{\prod_{n\in\mathcal{N}_k}(1+\gamma_{k,n})^{\frac{-\beta_k}{N}}\right\}, \end{aligned} \tag{4}$$

where $\beta_k=\frac{\theta_k B}{\ln(2)}$ with θ as delay exponent and $\mathbb{E}_\gamma$ gives the statistical average of inner arguments with respect to channel g. Each user selects its suitable RRHs concerning its tolerable delay parameter(θ) and the data queue length of RRHs. It is noteworthy that CoMP selection affects the amount of E_c. Each RRH can be considered in the CoMP selection, if and only if the arrival rate of the incoming traffic to that RRH is smaller than the achievable E_c. If this constraint is not satisfied, that RRH will not be selected in the RRH clustering. Hence, the E_c must be larger than the arrival rate which is equal to the fronthaul average rate in our system model, to guarantee the queuing stability at each RRH and to avoid any data loss. Hence, this condition is represented by

$$\text{C4:}\quad E_c^k\geq R_m^{\text{ave-max}},\ \forall k,m\in\ \mathcal{K},\mathcal{M}. \tag{5}$$

where $R_m^{\text{ave-max}}$ is the average fronthaul rate at RRH m. Furthermore, the maximum total available transmit power of each RRH is limited, i.e.,

$$\text{C5:}\;\; \mathbb{E}_g\left\{\sum_{n=1}^{N}\sum_{k=1}^{K} p_{k,m,n}\right\} \leq P_m^{\max},\;\forall m\in\mathcal{M}. \tag{6}$$

where $P_m^{\max}$ is the maximum power of RRH m. Next, we consider a limitation on fiber fronthaul rate for RRH m. At each RRH m, the average rate of the wireless access needs to be smaller than the average fronthaul rate of fiber link, i.e.,

$$\text{C6:}\sum_{n=1}^{N}\alpha_{m,n}\sum_{k=1}^{K}\phi_{k,n}\mathbb{E}_g\left\{\frac{B}{N}\log_2(1+\gamma_{k,n})\right\}\leq R_m^{\text{ave-max}},\forall m\in\mathcal{M}. \tag{7}$$

In this paper, our objective is to tackle the radio resource allocation and CoMP selection problem to guarantee the users' delay requirements in user-centric OFDMA-based C-RAN architecture. Hence, it is formulated as

$$\max_{\{p_{k,m,n},\alpha_{m,n},\phi_{k,n}\}} \sum_{k=1}^{K} E_c^k \tag{8}$$

Subject to: C1-C6.

Remark 1: It is worth noting that the proposed problem is not trivial because due to the expectation operator in the effective capacity formulation, the EC maximization problem is technically more challenging than the Shannon capacity maximization. Moreover, it defines a novel framework to jointly CoMP selection and resource allocation based on the statistical delay requirements to guarantee the delay in the user-centric C-RAN scenario.

III. Proposed Solution

The joint optimization problem is a mixed-integer nonlinear problem (MINLP) due to the integer variables $\{\alpha_{m,n},\phi_{k,n}\}$, which is prohibitively difficult to solve. Therefore, an iterative-algorithm is derived through converting the initial MINLP into a two-step optimization framework. In the first step, subcarrier assignment and power allocation are carried out for the fixed CoMP selection while in the second step, CoMP selection is updated by proposed heuristic algorithm. Each of these two subproblems can be solved with less complexity, which leads to a computationally convenient solution. However, the first subproblem is not jointly convex regarding power and subcarriers. Based on several previous works on multicarrier systems [19], [20], one approach to deal with the integer variable $\{\phi_{k,n}\}$ is relaxation approach, i.e. $\{\phi_{k,n}\}\in[0\;\;1]$ for all n and k. Then, the problem becomes jointly convex and it can be solved with a duality approach. By applying this technique, the Lagrangian function is

$$\begin{aligned}&\mathcal{L}(p_{k,m,n},\phi_{k,n},\omega_n,\lambda_{k,m},\mu_m,\zeta_m)=\sum_{k=1}^{K}E_c^k-\sum_{n=1}^{N}\omega_n\Big(\sum_{k=1}^{K}\phi_{k,n}-1\Big)\\&+\sum_{k=1}^{K}\sum_{m=1}^{M}\lambda_{k,m}\Big(E_c^k-R_m^{\text{ave-max}}\Big)+\sum_{m=1}^{M}\mu_m\Big(P_{\max}^m-\mathbb{E}_g\Big\{\sum_{k=1}^{K}\sum_{n=1}^{N}p_{k,m,n}\Big\}\Big)\\&+\sum_{m=1}^{M}\zeta_m\Big(R_m^{\text{ave-max}}-\sum_{n=1}^{N}\alpha_{m,n}\sum_{k=1}^{K}\phi_{k,n}\mathbb{E}_g\Big\{\frac{B}{N}\log_2(1+\gamma_{k,n})\Big\}\Big).\end{aligned} \tag{9}$$

The Lagrangian function, (9) is decomposed into a slave and master problem. The slave problem maximizes the transmit powers and assigns subcarriers for fixed Lagrange multipliers. The master problem updates the Lagrange multipliers by using the gradient-based method.

Lemma1: For a given set of Lagrange multipliers, the optimal transmit power allocation and subcarrier assignment are given as

$$\begin{aligned}p_{k,m,n}=[&\frac{\sigma^2}{\mathcal{K}_{k,m,n}\alpha_{m,n}|g_{k,m,n}|^2}\prod_{n\in\mathcal{N}_k}\Big(\frac{1}{\mathcal{K}_{k,m,n}}\Big)^{\left(\frac{-\beta_k}{N(1+\beta_k)}\right)}\\&-\frac{\sigma^2(1+\sum_{i=1,i\neq k}^{M}\alpha_{i,n}|g_{k,i,n}|^2p_{k,i,n})}{\alpha_{m,n}|g_{k,m,n}|^2}]^+,\end{aligned} \tag{10}$$

where $\mathcal{K}_{k,m,n}=\frac{N\ln(2)\mu_m\sigma^2\delta_2+\sum_{i=1}^{M}\zeta_i\alpha_{i,n}\phi_{k,n}B\alpha_{m,n}|g_{k,m,n}|^2}{\delta_2(1+\lambda_{k,m})B\alpha_{m,n}|g_{k,m,n}|^2(\ln 2\delta_1)^{(-1)}}$ and $\delta_1=\mathbb{E}_g\Big\{\prod_{n\in\mathcal{N}_k}(1+\gamma_{k,n})^{-\frac{\beta_k}{N}}\Big\}$ and $\delta_2=\mathbb{E}_g\Big\{\frac{B}{N}(1+\gamma_{k,n})\Big\}$. Moreover,

$$\phi_{k,n}=\begin{cases}1 & \text{If } k=\max_i\Psi_{n,i},\\ 0 & \text{otherwise.}\end{cases} \tag{11}$$

where $\Psi_{n,}$ is introduced as the marginal benefit of assigning the subcarrier n to the user .

Proof: See Appendix A ■

The next step is to obtain the CoMP selection strategy for fixed $\{p_{k,m,n},\phi_{k,n}\}$. The optimal solution requires an exhaustive search algorithm to search for all possible CoMP selections and users. Hence, the overall complexity is $\mathcal{O}(2^M NK)$, which is impractical with a large number of M. To overcome this burden, a heuristic algorithm with lower complexity is introduced. In order to apply this algorithm, let us define $\hat{\gamma}_{k,n}(A_n)$ as the SNR of user k over the selected set of RRHs A_n, which is defined as $\hat{\gamma}_{k,n}(A_n)=\frac{\sum_{m\in A_n}\alpha_{m,n}p_{k,m,n}|g_{k,m,n}|^2}{\sigma^2}$. Afterward, the Lagrangian function over the selected set of RRHs A_n is given by

$$\begin{aligned}&\mathcal{L}_n(p_{k,m,n},A_n)=\sum_{k=1}^{K}\hat{E}_c^k+\sum_{k=1}^{K}\sum_{m\in A_n}\lambda_{k,m}\Big(\hat{E}_c^k-R_m^{\text{ave-max}}\Big)+\sum_{m\in A_n}\mu_m\\&\Big(P_{\max}^m-\mathbb{E}_g\Big\{\sum_{k=1}^{K}\sum_{n=1}^{N}p_{k,m,n}\Big\}\Big)+\sum_{m\in A_n}\zeta_m\Big(R_m^{\text{ave-max}}-\sum_{n=1}^{N}\alpha_{m,n}\sum_{k=1}^{K}\phi_{k,n}\\&\mathbb{E}_g\Big\{\frac{B}{N}\log_2(1+\hat{\gamma}_{k,n})\Big\}\Big),\end{aligned} \tag{12}$$

where $\hat{E}_c^k=\frac{-1}{\theta_k}\log\mathbb{E}_g\Big\{\prod_{n\in\mathcal{N}_k}(1+\hat{\gamma}_{k,n}(A_n))^{\frac{-\beta_k}{N}}\Big\}$. The objective function $F(A_n)$ for the heuristic algorithm is then defined as $F(A_n)=\max_{A_n}\mathcal{L}_n(p_{k,m,n},A_n)$, Since $p_{k,m,n}$ is fixed, each user k tends to associate with a subset of RRHs satisfying this objective function. It means that RRH m is selected if it satisfies in the following equation,

$$F(A_n\cup\{m\})>F(A_n). \tag{13}$$

Consequently, we achieve the CoMP selection in M iterations as follows. We define F^j as the objective value of the iteration j. At first, we initiate $A_n=\emptyset$ and $F^j=0$. Moreover, the channel power gains on each SC n are sorted descending. At each iteration $j\in\{1,...,M\}$, we select the RRH m_j with

Algorithm I: Proposed Heuristic Algorithm

I)Randomly initialize $\alpha_{m,n}, \lambda_{k,m}, \mu_m, \zeta_m, \omega_n$ and $\epsilon > 0$
while $\{\lambda_{k,m}, \mu_m, \zeta_m, \omega_n\}$ *converge* **do**
  repeat
    II) Obtain $p_{k,m,n}$ and $\phi_{k,n}$ by (10) and (11) with given $\alpha_{m,n}$;
    Initialize $A_n^1 = \emptyset$, $F^1 = 0$ and sort $g_{k,m,n}$ descendly;
    for *j=1:M* **do**
      Select RRH m_j according to the sorted channel;
      if m_j *satisfies in* (13) **then**
        a: Update selected RRH set as $A_n = A_n \cup \{m_j\}$;
        b: Update objective F^{j+1} ;
      else
        c: Remains A_n unchanged and $F^{j+1} = F^j$;
      end
    end
    III) Update $\alpha_{m,n}$ with given $p_{k,m,n}$ and $\phi_{k,n}$.
  until; {Convergence within the tolerance ϵ.}
  IV) Update Lagrange multipliers $\lambda_{k,m}, \mu_m, \zeta_m, \omega_n$,
end

the j-th largest channel power gain. Afterwards, the equation (13) is checked. If it improves the value of F^j, the RRH m_j is added to the subset of A_n. Consequently the set A_n and the objective value F^{j+1} are updated as $A_n = A_n \cup \{m_j\}$ and $F^{j+1} = F(A_n \cup \{m_j\})$ respectively. After M iterations, the final RRH subset A_n on SC n is obtained. It is noticeable that based on 13, we can not guarantee if $p_{k,m,n} \neq 0$ then $\alpha_{m,n} = 1$. Moreover, the power allocation depends on the entire CoMP selection A_n and even if $p_{k,m,n} \neq 0$, it can not be known beforehand whether this particular CoMP selection $\alpha_{m,n}$ is the one that maximizes the objective of the problem 13. In this heuristic algorithm, channel sorting has the computational complexity of $\mathcal{O}(M \log_2(M))$ and the complexity of M iterations is $\mathcal{O}(M)$. Since they are independent, the overall complexity of the Problem (8) is $\mathcal{O}(M^2 \log_2(M)KN)$, which is much lower than the exhaustive search algorithm especially for the large number of M. The proposed algorithm is described in detail in Algorithm I. It is worth noting that the CoMP selection is performed based on the channel gains and the delay requirements as it is shown in Algorithm I. Firstly, the RRHs are selected according to the channel gains, and then the equation (13) is checked to investigate the delay requirements satisfaction.

IV. Simulation Results

For the simulation setup in the user-centric OFDMA-based 5G C-RAN, we focus on a scenario of a cluster of M RRHs serving K users, and both of them are uniformly distributed in the whole area of interest. We consider channels with Rayleigh fading distribution model. Specifically, the channel is modeled as $\alpha^2|\frac{d_0}{d_k}|^{(2.5)}$ where α is a standard complex circularly symmetric Gaussian random variable models fading effects and $|\frac{d_0}{d_k}|$ models the communication power path-loss at reference distance denoted by $d_0 = 1m$ and the distance between the transmitter and the user k denoted by d_k. The average rate of the each fronthaul link is $R_m^{\text{ave-max}} = 1$ Mbps. The wireless access channel has a total bandwidth of 20 MHz which is divided into 128 SCs. It is noted that our proposed problem is the first framework to jointly allocate the resources and CoMP selection by considering the statistical delay. Hence, for the performance comparisons to show the benefits gathered from

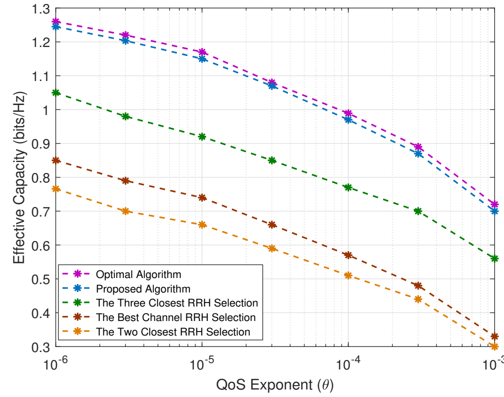

Figure 2: E_c versus delay exponent θ for 10 users, 6 RRHs

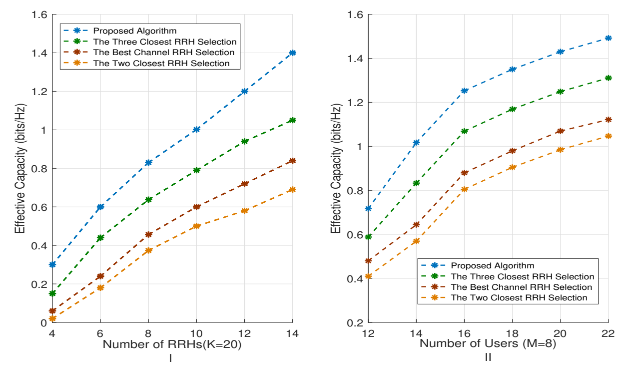

Figure 3: E_c for delay exponent $\theta = 10^{-5}$;I) E_c versus number of RRHs, II) E_c versus number of users

our proposed algorithm, we consider two scenarios. The first one is that the users can associate with the closest L RRHs. Here, we consider both the two and three closest RRHs. The other scheme is to select the one RRH with the best channel gain. Fig. 2 expresses the delay behavior of the proposed algorithm and the benchmarks. It is shown that the E_c attained by all schemes decreases with the delay exponent (θ). Hence, there is a trade-off between delay and achievable throughput. It means that when the delay constraint is tight, the E_c is much lower than the looser delay requirements. Note that as θ goes to zero, the E_c equals Shannon capacity, and the delay requirements do not play an important role. It is observed that our proposed heuristic solution has a much better performance than the other benchmarks. It means that we can achieve higher throughput by applying the proposed CoMP selection strategy. Furthermore, it reveals that the proposed heuristic solution performs almost as well as the optimal solution with a negligible gap. Fig. 3 illustrates the variation of the E_c regarding the number of RRHs and users. It is clear that the E_c is an ascending function of the number of RRHs and users. Moreover, it shows that the heuristic algorithm outperforms the other scenarios, which implies the benefits of the proposed user-centric CoMP selection algorithm which effectively allocates the resources. Also, the benchmarks comparisons show that not only cooperation but also the optimal RRH channel selection have impacts on the throughput.

V. Conclusion

In this paper, we study downlink transmission of 5G OFDMA-based user-centric C-RAN by considering the delay requirement guarantees. A novel framework based on an effective capacity notion is proposed to optimize the problem of subcarrier-user assignment, CoMP selection, and power allocation subject to delay and power constraints.An efficient solution is proposed based on duality and greedy algorithms for the combinatorial and non-convex problem. Numerical simulation results illustrate that the heuristic algorithm achieves the very near-optimal throughput performance and the proposed algorithm significantly outperforms the other benchmarks, and increases the spectral efficiency.

APPENDIX

In order to solve the optimal power allocation $p_{k,m,n}$, we introduce an additional index l for each allocated power to show L fading states (or sub-channels). For notational ease, we introduce $\mathcal{F}_n = (1+\gamma_{k,n})^{\frac{-\beta_k}{N}}$. With this definition, the first order derivative of $\mathcal{L}$ with respect to $p_{k,m,n,l}$ can be written as,

$$\frac{\partial \mathcal{L}}{\partial p_{k,m,n,l}} = \frac{-(1+\lambda_{k,m})\beta_k |g_{k,m,n,l}|^2}{\theta_k \ln 2\bar{\delta_1} N\sigma^2}(1+\gamma_{k,n,l})^{(\frac{-\beta_k}{N}-1)}$$

$$\prod_{j\in\mathcal{N}_k, j\neq n} \mathcal{F}_j - \mu_m - \sum_{i=1}^{M} \zeta_i \alpha_{i,n}\phi_{k,n} \frac{B\alpha_{m,n}|g_{k,m,n,l}|^2}{N\ln 2\sigma^2 \bar{\delta_2}} = 0, \tag{14}$$

where $\bar{\delta_1} = \frac{1}{L}\sum_{l=1}^{L}\prod_{n\in\mathcal{N}_k}(1+\gamma_{k,n,l})^{-\frac{\beta_k}{N}}$ and $\bar{\delta_2} = \frac{1}{L}\sum_{l=1}^{L}\left\{\frac{B}{N}(1+\gamma_{k,n,l})\right\}$. Then, Eq. (14) can be simply rewritten as $(1+\gamma_{k,n,l})^{-1}\prod_{n\in\mathcal{N}_k}\mathcal{F}_n = \mathcal{K}_{k,m,n,l}$, where $\mathcal{K}_{k,m,n,l}$ is the cut-off threshold and denoted by $\mathcal{K}_{k,m,n,l} = \frac{N\ln(2)\mu_m\sigma^2\bar{\delta_2}+\sum_{i=1}^{M}\zeta_i\alpha_{i,n}\phi_{k,n}B\alpha_{m,n}|g_{k,m,n}|^2}{\bar{\delta_2}(1+\lambda_{k,m})B\alpha_{m,n}|g_{k,m,n}|^2(\ln 2\bar{\delta_1})^{(-1)}}$. Hence, we can express $\gamma_{k,n,l}$ in terms of $\gamma_{k,j,l}$ as $(1+\gamma_{k,j,l}) = \frac{\mathcal{K}_{k,m,n,l}}{\mathcal{K}_{k,m,j,l}}(1+\gamma_{k,n,l})$. Consequently, we have

$$\prod_{n\in\mathcal{N}_k}\mathcal{F}_n = (1+\gamma_{k,n,l})^{-\beta_k}\prod_{j\in\mathcal{N}_k, j\neq n}\left(\frac{\mathcal{K}_{k,m,n,l}}{\mathcal{K}_{k,m,j,l}}\right)^{\frac{-\beta_k}{N}}. \tag{15}$$

Therefore after some calculation, we finally obtain

$$p_{k,m,n,l} = \Big[\frac{\sigma^2}{\mathcal{K}_{k,m,n,l}\alpha_{m,n}|g_{k,m,n,l}|^2}\prod_{n\in\mathcal{N}_k}\left(\frac{1}{\mathcal{K}_{k,m,n,l}}\right)^{(\frac{-\beta_k}{N(1+\beta_k)})}$$
$$- \frac{\sigma^2(1+\sum_{i=1,i\neq k}^{M}\alpha_{i,n}|g_{k,i,n,l}|^2 p_{k,i,n,l})}{\alpha_{m,n}|g_{k,m,n,l}|^2}\Big]^+. \tag{16}$$

Similar to [21], we let $L\to\infty$, which yields to Eq. (10) while $\bar{\delta_1}\to\delta_1$ and $\bar{\delta_2}\to\delta_2$. Next, to find the optimal subcarrier assignment, the first derivative of the Lagrangian function $\mathcal{L}$ with respect to $\phi_{k,n}$ for a given optimal power allocation $p_{k,m,n}$ is as follows,

$$\frac{\partial\mathcal{L}}{\partial\phi_{k,n}} = \frac{(1+\lambda_{k,m})\log(1+\gamma_{k,n})e^{\left(-\frac{\theta_k B}{N}\sum_{n=1}^{N}\phi_{k,n}\log(1+\gamma_{k,n})\right)}}{N\ln 2\mathbb{E}_g\left\{\prod_{n\in\mathcal{N}_k}(1+\gamma_{k,n})^{\frac{-\beta_k}{N}}\right\}}$$
$$-\omega_n - \sum_{i=1}^{M}\zeta_i\alpha_{i,n}\mathbb{E}_g\left\{\frac{B}{N}(1+\gamma_{k,n})\right\} \triangleq \Psi_{k,n}, \tag{17}$$

$\Psi_{k,n}$ is introduced as the marginal benefit of assigning the subcarrier n to the user k. Hence, there are a set $\{\Psi_{1,n}, ..., \Psi_{K,n}\}$ with K items relating the subcarrier n to the different users. Therefore, the user i with maximum marginal benefit, $\Psi_{i,n}$ will be assigned to subcarrier n as in equation (11).

REFERENCES

[1] M. Sinaie, P.-H. Lin, A. Zappone, P. Azmi, and E. A. Jorswieck, "Delay-aware resource allocation for 5G wireless networks with wireless power transfer," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 7, pp. 5841–5855, 2018.

[2] L. Ferdouse, A. Anpalagan, and S. Erkucuk, "Joint communication and computing resource allocation in 5G cloud radio access networks," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 9, pp. 9122–9135, 2019.

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

[4] Z. Lin and Y. Liu, "Joint uplink and downlink transmissions in user-centric OFDMA Cloud-RAN," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 8, pp. 7776–7788, 2019.

[5] C. Pan, H. Ren, M. Elkashlan, A. Nallanathan, and L. Hanzo, "Weighted sum-rate maximization for the ultra-dense user-centric TDD C-RAN downlink relying on imperfect CSI," *IEEE Transactions on Wireless Communications*, vol. 18, no. 2, pp. 1182–1198, 2019.

[6] L. You and D. Yuan, "User-centric performance optimization with remote radio head cooperation in C-RAN," *IEEE Transactions on Wireless Communications*, vol. 19, no. 1, pp. 340–353, 2019.

[7] *Study on New Radio Access Technology: Radio Access Architecture and Interfaces*, 3GPP document 38.801 NTT Docomo 2017.

[8] *Technical Specification Group Radio Access Network: NG-RAN Architecture Description*, 3GPP document TS 38.401 V15.6.0 Rel-15 2019.

[9] A. Alabbasi, X. Wang, and C. Cavdar, "Optimal processing allocation to minimize energy and bandwidth consumption in hybrid CRAN," *IEEE Transactions on Green Communications and Networking*, vol. 2, no. 2, pp. 545–555, 2018.

[10] Y. Luo, K. Yang, Q. Tang, J. Zhang, P. Li, and S. Qiu, "An optimal data service providing framework in cloud radio access network," *EURASIP Journal on Wireless Communications and Networking*, no. 1, pp. 1–11, 2016.

[11] Z. Zhao, M. Peng, Z. Ding, W. Wang, and H. V. Poor, "Cluster content caching: An energy-efficient approach to improve quality of service in cloud radio access networks," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 5, pp. 1207–1221, 2016.

[12] J. Luo, Q. Chen, and L. Tang, "Reducing power consumption by joint sleeping strategy and power control in delay-aware C-RAN," *IEEE Access*, vol. 6, pp. 14655–14667, 2018.

[13] J. Li, M. Peng, A. Cheng, Y. Yu, and C. Wang, "Resource allocation optimization for delay-sensitive traffic in fronthaul constrained cloud radio access networks," *IEEE Systems Journal*, vol. 11, no. 4, pp. 2267–2278, 2014.

[14] D. Wu and R. Negi, "Effective capacity: a wireless link model for support of quality of service," *IEEE Transactions on wireless communications*, vol. 2, no. 4, pp. 630–643, 2003.

[15] D. W. K. Ng, E. S. Lo, and R. Schober, "Energy-efficient resource allocation in multiuser OFDM systems with wireless information and power transfer," in *IEEE Wireless communications and networking conference (WCNC)*, pp. 3823–3828, IEEE, 2013.

[16] J. Francis, J. K. Chaudhary, A. N. Barreto, and G. Fettweis, "Uplink latency in Massive MIMO-based C-RAN with intra-PHY functional split," *IEEE Communications Letters*, vol. 24, no. 4, pp. 912–916, 2020.

[17] H. Ren, N. Liu, C. Pan, M. Elkashlan, A. Nallanathan, X. You, and L. Hanzo, "Low-latency C-RAN: An next-generation wireless approach," *IEEE vehicular technology magazine*, vol. 13, no. 2, pp. 48–56, 2018.

[18] H. Ren, N. Liu, C. Pan, X. You, and L. Hanzo, "Power-and rate-adaptation improves the effective capacity of C-RAN for Nakagami-m fading channels," *IEEE Transactions on Vehicular Technology*, vol. 67, pp. 10841–10855, 2018.

[19] C. C. Zarakovitis and Q. Ni, "Maximizing energy efficiency in multiuser multicarrier broadband wireless systems: Convex relaxation and global optimization techniques," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 7, pp. 5275–5286, 2015.

[20] Y.-F. Liu and Y.-H. Dai, "On the complexity of joint subcarrier and power allocation for multi-user OFDMA systems," *IEEE transactions on Signal Processing*, vol. 62, no. 3, pp. 583–596, 2013.

[21] T. David and P. Viswanath, *Fundamentals of wireless communications*. Cambridge University Press, 2005.