IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 10, 2019, accepted December 25, 2019, date of publication January 24, 2020, date of current version February 11, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2969280

Flexible Function Splitting and Resource Allocation in C-RAN for Delay Critical Applications

MASUMEHSADAT TOHIDI[1], HAMIDREZA BAKHSHI[1], AND SAEEDEH PARSAEEFARD[2], (Senior Member, IEEE)
[1]Department of Electrical Engineering, Shahed University, Tehran 33191-18651, Iran
[2]Iran Telecommunications Research Center, Tehran 1439955471, Iran

Corresponding author: Hamidreza Bakhshi (bakhshi@shahed.ac.ir)

ABSTRACT The concept of cloud radio access network (C-RAN) architecture is being proposed to fully meet the requirements of 5G mobile networks. Thanks to the centralized cloud baseband unit (BBU), C-RAN reduces the energy consumption and cost of deployment significantly. However, it suffers from stringent fronthaul capacity and latency which are substantial in delay critical applications. Splitting up the processing functionalities between the control unit (CU) and distributed units (DUs) can mitigate fronthaul load and relax their requirements with the expense of an increase in power consumption. In this paper, we investigate joint access and fronthaul resource allocation problem for delay critical applications where the objective is minimizing the sum of normalized total power and fronthaul bandwidth consumption. We consider a downlink scenario and incorporate the total end-to-end delay components. For simplicity, linear models are assumed between function splitting (FS) levels and decreased fronhaul load as well as increased processing power. Different delay requirements affect our objective function and enforce a different FS level. We establish a flexible decision about the best FS level, which minimize our objective. Simulation results demonstrate that the delay constraint has a significant impact on the required fronthaul bandwidth and power consumption, which are directly related to the cost of the network. Moreover, flexible selecting function split level can achieve up to 40% gain in reducing the total utility (i.e., the sum of normalized total power and fronthaul required bandwidth).

INDEX TERMS C-RAN, resource allocation, delay critical, flexible function splitting.

I. INTRODUCTION

Cloud radio access network (C-RAN) is a prospective 5G wireless system architecture, which is anticipated to reduce capital expenditure (CAPEX) and operating expense (OPEX) and provide a higher quality of service (QoS) for users [1]. C-RAN consists of three main parts: remote radio head (RRH), fronthaul link, and baseband unit (BBU) pool. Centralizing baseband processing functionalities in BBU reduces the energy consumption of the system and facilitates expanding the scale of the network at a low cost. Moreover, a software-defined structure introduces flexibility in C-RAN to manage, upgrade the system and support the new functionalities [2].

Besides these advantages, C-RAN faces basic challenges in fronthaul link capacity and stringent delay requirement. To overcome these issues, different solutions are proposed [3]. Function splitting (FS) is a new strategy to reduce the fronthaul bit rate which is related to the software-defined feature of C-RAN. This capability enables operators to dynamically elect between centralized high processing powered datacenter and distributed antenna site for realizing functions module, based on different application scenarios [4], [5].

The 3rd generation partnership project (3GPP) protocol stack, introduces eight function split levels, as being depicted in Fig. 1 [6]. In this protocol stack, the baseband functions, divided into the distributed unit (DU) and centralized unit (CU). In Fig. 1, the functions located on the right side of the split levels separator, are accomplished in CU, and the left of it are done in DU. In split level 1, only RF functions

The associate editor coordinating the review of this manuscript and approving it for publication was Adnan Shahid.

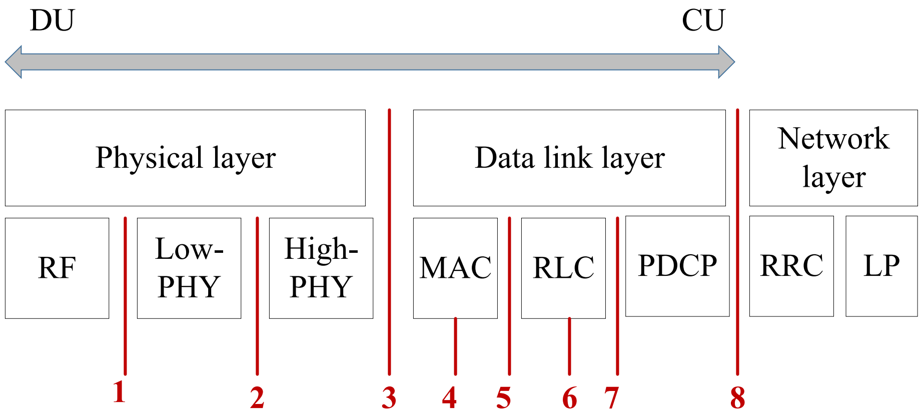

FIGURE 1. Function splitting level proposed by 3GPP [6].

are implemented in DU and have been known as traditional C-RAN. On the other hand, at split 8, all data link and physical layer functions are executed in DU. In general, lower split levels, load more fronthaul data, which fulfills the rate requirement. For the higher split level when the more functions are performed in DU, the lower fronthaul capacity is required [6]–[8].

In traditional C-RAN, due to the centralized processing, power consumption of the system is greatly reduced. Transferring all processing functions to the CU requires high-capacity and low-delay fronthaul links [9]. As the more functions are locally implemented in DU, the required processing power and therefore the total cost of the system are increased. Given the mentioned trade-off, flexible processing function splitting according to the channel variations and users' requirements is necessary. This concept is recently discussed in [10]–[13]. In [14], a flexible RAN architecture with a user centric level of functional splitting is investigated. Authors address the new orchestration of heterogeneous resources in a multi-sited C-RAN infrastructure to optimize jointly the functional split and End-to-End resource allocation in order to achieve an enhanced throughput satisfaction and a low deployment cost.

Another important parameter that varies with FS levels is delay. A large body of works has been investigated the effect of FS on the total delay of the system. Taking into account the worst-case delay model, the authors in [15] have studied its suitability in defining dimensioning rules for a number of cellular scenarios, where the fronthaul traffic flows follow splits I_U and II_D in eCPRI specification. They use a G/G/I queueing model as the key metric in fronthaul network dimensioning for split I_U and with numerical results, Kingman's estimation for different link rates are computed. The tradeoff between fronthaul bitrate, flexibility, and complexity of the local equipment is investigated in [16]. This work demonstrates the advantages of using a packet-switched fronthaul with different functional split options for delay constrained systems. Authors in [17], by mathematical and simulation methods, evaluate different splits with respect to network energy level and cost efficiency for expected delay quality of service and derive a principle for fronthaul dimensioning based on the traffic profile. In [18], an analytical framework to calculate the latency in the uplink of C-RAN massive MIMO system with functional split 7 is presented. They model the output port of an Ethernet switch as an M/HE/1 queue and derive closed-form expressions for sojourn time and queue length distribution.

The effect of delay constraint on C-RAN with flexible function splitting is discussed in some recent works. In [19], the impact of different split points on the system's energy and mid-haul link bandwidth consumptions are investigated. The authors propose an optimization framework to incorporate the end-to-end delay, from the central cloud to the end-user, under different/ flexible FS points. Here, the delay is defined as the delay induced by processing at the central cloud and/or edge-cloud, mid-haul and fronthaul transportation delay, and radio access transmission delay. A flexible designed structure for resource allocation in density aware C-RAN is studied in [3]. The authors consider two design mode RAN based on the average density of users. In high-density regime, the rate and power consumption are considered as two major conflicting objective functions. The sum of the transmission delay and cost of processing in DUs are minimized in the low-density region where the level of splitting is confided as the variables of the optimization problem. Authors of [11], propose a modeling approach and a rigorous analytical framework, FluidRAN, for minimizing RAN costs by jointly selecting the splits under delay constraint. The impact of different functional splits on the FH capacity and latency is studied in [20]. Focusing on uRLLC, eMBB and mMTC, in [21] authors analyze how low-latency and high bandwidth requirements of these traffic classes are met by providing different split between CU and DU. They investigate the optimal placement for the service function chains based on optimization goals for different network slices.

In the aforementioned works in the preceding paragraph, the problem of resource allocation with flexible FS for delay critical application was not investigated. Our contributions are summarized as follows:

- We formulate a joint fronthaul and access resource allocation problem for a downlink C-RAN based architecture which can flexibly split up functions into DUs and CU while tacking into account limited power and delay constraints. We consider a wireless fronthaul link and our main objective is to minimize the normalized system power and required bandwidth. To save power, DUs have the ability to be off or on, and the number of virtual machines (VM) use for process in CU can adaptively be determined by the total demand rate of it.
- To tackle the delay constraint, both processing and transportation delay (sum of transmitting and queueing delay) are tacking into account. A linear relationship between processing delay and FS level is assumed. We present a practical double-layer queueing model where the first layer is for aggregated transmit data of each user and the second is for data of each user.
- To overcome the non-convex nature of the proposed problem, we utilize an iterative algorithm based on the BCD method for all eight FS levels separately, then select the best level which minimizes our objective function. Simulation results demonstrate an interconnection between delay constraint and required fronthaul bandwidth as well as consumed power of the system, hence enforced FS level.

The rest of this paper is structured as follows. The system and queueing models are detailed in Sec. II. The optimization problem formulation and the proposed algorithm are presented in Sec. III. The complexity and convergence analysis of our solution is discussed in Sec. IV. The simulation results are reported in Sec. V. Finally, Sec. VI concludes the paper.

II. SYSTEM MODEL

In this section, we present our considered setup, queueing model, and practical constraints.

A. SYSTEM DESCRIPTION

Consider a partially centralized multi-cell C-RAN in downlink transmission mode. In this setup, the radio processing and baseband functions from the 3GPP protocol stack are flexibly split up into eight levels between DUs and CU as shown in Fig. 1. We define the variable L for determining the FS level as

$$\mathrm{C1}: L \in \{1, 2, \ldots, 8\}.$$

In Fig. 1, the red line indicates the function split level where the left functions of it will be implemented in DU and right of it will be done at CU. We consider $\mathcal{M} = \{1, 2, \ldots, M\}$ DUs which are served to $\mathcal{K} = \{1, 2, \ldots, K\}$ users. To save power, each DU m can be switched off, which corresponds to the sleep mode. This capability express by the variable b_m which define as follows

$$b_m = \begin{cases} 1, & \text{if DU } m \text{ is ON,} \\ 0, & \text{if DU } m \text{ is OFF.} \end{cases} \tag{1}$$

CU includes a set of $\mathcal{B} = \{1, 2, \ldots, B\}$ VMs which operate as virtual base stations to process baseband signals and optimize the network resource allocation tasks. We assume that this part connects to DUs with a wireless fronthaul link. In Fig. 2, the overall scheme of our considered system model is plotted.

B. ACCESS LINK MODEL AND PARAMETERS

We consider OFDMA between users where W_{A} is the available bandwidth of each sub-carriers for access and the set of them is indicated by $\mathcal{N}_{\mathrm{A}} = \{1, 2, .., N_{\mathrm{A}}\}$. The channel vector between user k, DU m, and sub-carrier n_{A} is denoted by $h_{k,n_{\mathrm{A}},m}$. We introduce the binary variable $\alpha_{k,n_{\mathrm{A}},m}$, which specifies sub-carrier allocation for each user as

$$\alpha_{k,n_{\mathrm{A}},m} = \begin{cases} 1, & \text{if the sub-carrier } n_{\mathrm{A}} \text{ is assigned to user } k, \\ 0, & \text{else.} \end{cases} \tag{2}$$

Due to OFDMA limitation, each sub-carrier should be assigned to only one user, therefore,

$$\mathrm{C2}: \sum_{k \in \mathcal{K}} \alpha_{k,n_{\mathrm{A}},m} \leq 1, \quad \forall n_{\mathrm{A}} \in \mathcal{N}_{\mathrm{A}}, \ \forall m \in \mathcal{M}.$$

Let $\beta_{k,m}$ indicate whether user k is assigned to DU m, i.e.,

$$\beta_{k,m} = \begin{cases} 1, & \text{if the user } k \text{ is assigned to DU } m, \\ 0, & \text{else.} \end{cases} \tag{3}$$

We assume that each user is connected to only one DU which leads to

$$\mathrm{C3}: \sum_{m \in \mathcal{M}} \beta_{k,m} \leq 1, \quad \forall k \in \mathcal{K}.$$

When at least one user is assigned to a DU, this DU is active, therefore we have the following constraint.

$$\mathrm{C4}: \beta_{k,m} \leq b_m, \quad \forall k \in \mathcal{K}, \ \forall m \in \mathcal{M}.$$

The total achievable rate of user k is equal to

$$R_k = \sum_{m \in \mathcal{M}} \sum_{n_{\mathrm{A}} \in \mathcal{N}_{\mathrm{A}}} W_{\mathrm{A}} \beta_{k,m} \alpha_{k,n_{\mathrm{A}},m} \log(1 + \frac{p_{k,n_{\mathrm{A}},m} h_{k,n_{\mathrm{A}},m}}{\sigma^2 + I_{k,n_{\mathrm{A}},m}}), \quad \forall k \in \mathcal{K}, \tag{4}$$

where $p_{k,n_{\mathrm{A}},m}$ represent the transmit power, σ^2 is the noise power at the receiver of user k and $I_{k,n_{\mathrm{A}},m} = \sum_{\substack{m' \in \mathcal{M},\\ m' \neq m}} \sum_{\substack{k' \in \mathcal{K}\\ k' \neq K}} p_{k',n_{\mathrm{A}},m'} h_{k,n_{\mathrm{A}},m'}$ is the intra-cell interference (between CUs). The maximum total available transmit power of each DU is limited, i.e.,

$$\mathrm{C5}: \sum_{k \in \mathcal{K}} \sum_{n_{\mathrm{A}} \in \mathcal{N}_{\mathrm{A}}} p_{k,n_{\mathrm{A}},m} \leq P_{\text{max-DU}}, \quad \forall m \in \mathcal{M},$$

where $P_{\text{max-DU}}$ is the maximum transmit power of each DU.

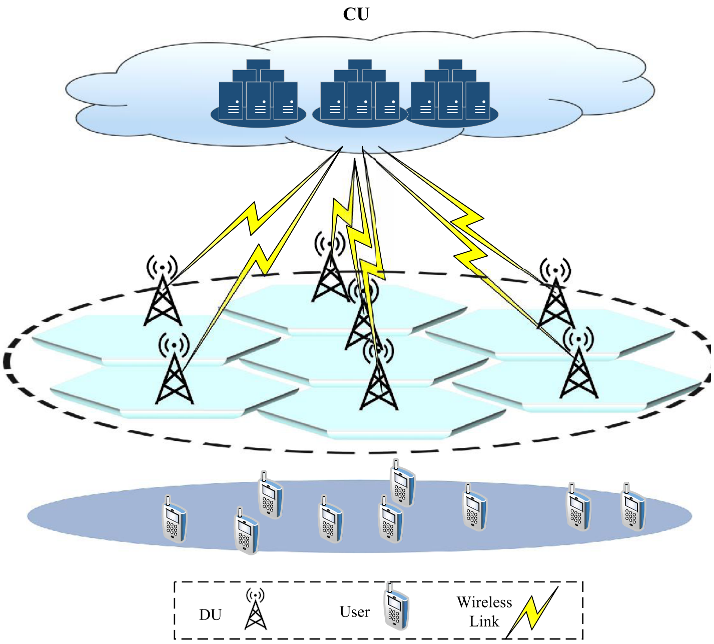

FIGURE 2. Considered system model. It is noticeable that both CU and DUs, have the capability of performing processing functions, regardless of selected FS level.

C. FRONTHAUL LINK MODEL AND PARAMETERS

Although optical fiber links can provide high transmission capacity, they have the high-cost expense and inflexible deployment. Wireless fronthaul links are cheap and flexible but are capacity-constrained and need to appropriate actions for dealing with interference and security [22]. In this paper, we consider wireless fronthaul link, where share between DUs through OFDMA with $\mathcal{N}_{\text{F}} = \{1, 2, \ldots, N_{\text{F}}\}$ as the set of sub-carriers and each has W_{F} bandwidth. The frequency bands of fronthaul links are considered to be separate from access links. The sub-carriers which are associated with DU m determine by a binary variable $\tau_{m,n_{\text{F}}}$ which defines as

$$\tau_{m,n_{\text{F}}} = \begin{cases} 1, & \text{if the sub-carrier } n_{\text{F}} \text{ is assigned to DU } m, \\ 0, & \text{else.} \end{cases} \tag{5}$$

Due to OFDMA, each sub-carrier can be assigned to at most one DU, i.e.,

$$\text{C6}: \sum_{m\in\mathcal{M}} \tau_{m,n_{\text{F}}} \leq 1, \quad \forall n_{\text{F}} \in \mathcal{N}_{\text{F}}.$$

The total achievable rate of DU m can be calculated as

$$r_m = \sum_{n_{\text{F}}\in\mathcal{N}_{\text{F}}} W_{\text{F}}\tau_{m,n_{\text{F}}} \log(1 + \frac{p_{m,n_{\text{F}}} g_{m,n_{\text{F}}}}{\sigma^2}), \tag{6}$$

where $g_{m,n_{\text{F}}}$ is the channel gain between DU m and CU in sub-carrier n_{F}, and $p_{m,n_{\text{F}}}$ is the transmit power of DU m to CU. Due to the power limitation of CU, we have,

$$\text{C7}: \sum_{n_{\text{F}}\in\mathcal{N}_{\text{F}}} p_{m,n_{\text{F}}} \leq P_{\text{max-CU}}, \quad \forall m \in \mathcal{M},$$

where $P_{\text{max-CU}}$ is the maximum transmit power of CU to each DU.

D. DELAY ANALYSIS AND QUEUING MODEL

Overall delay of the considered system comes from three components [24]–[26],

$$D_{\text{Total}} = D_{\text{Processing}} + D_{\text{Propagation}} + D_{\text{Transsport}}. \tag{7}$$

1) $D_{\text{Processing}}$: In the considered system model, the processing functions are done in CU or DUs, according to the selected FS level. Due to the strong processors in the CU, we neglect the processing delay of it and only consider the processing delay in the DU which is related to the FS level. At lower FS levels, more processing is done in the DUs so processing delay becomes longer. For simplicity, regardless of function type, we assume a linear relation between FS level and processing delay in DUs. Therefore, the processing delay of each user can be expressed by

$$D_{\text{Processing}} = \sum_{m\in\mathcal{M}} \beta_{k,m} L d_{\text{p}}, \tag{8}$$

where d_{p} is the normalized time for process in DUs.

2) $D_{\text{Propagation}}$: The propagation time is proportional to the physical distance between the transmitter and the receiver. For fixed DUs and CU, $D_{\text{Propagation}}$ is not varied and is a constant.

3) $D_{\text{Transport}}$: It consists of the time to radio transmission and queuing of the data. Queuing delay is due to temporary buffering data before transmission on the wireless fading channels. Determining an appropriate queuing model based on the considered setup is essential in calculating queueing delay.

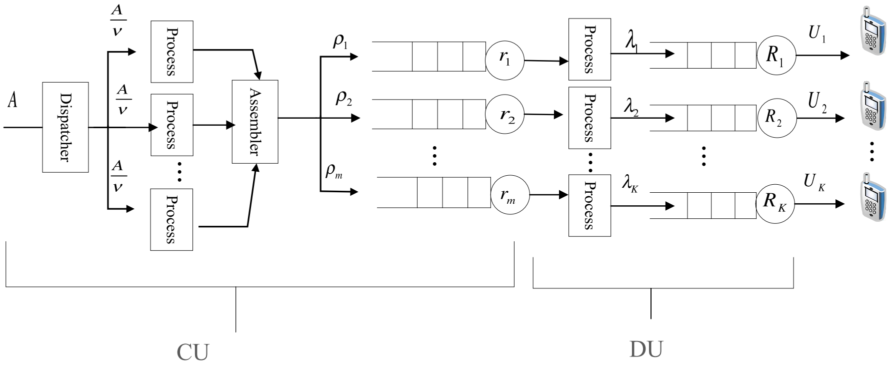

FIGURE 3. General E2E queuing model for C-RAN with FS.

We consider the arrival data of user's transmission queues having a Poisson distribution with mean arrival rate λ_k. Total user's data, i.e., $A = \sum_{k\in\mathcal{K}} \lambda_k$, is divided into ν VMs to process in CU. After processing, the data to be sent to each DU, place in a separate transmission queue. The considered E2E queueing model is illustrated in Fig.3. We note that this model is more practical in contrast to our previous work [27] in which we assume that each user data place in septate processing and transmission queues.

Depending on the levels of FS, the arrival bit rate of fronthaul transmission queue is varied. For simplicity, we assume that fronthaul bit rate ρ_m is a linear function of L. Consequentially, we have the following constraint:

$$\rho_m = L_0 L \sum_{k\in\mathcal{K}} \beta_{k,m}\lambda_k, \quad \forall m \in \mathcal{M}, \tag{9}$$

where L_0 is a constant. We suppose that the service time of fronthaul and access transmission queues follow exponential distributions with mean $1/r_m$ and $1/R_k$, respectively. Therefore we have two layers of M/M/I queues in tandem and the total queuing delay plus transmission delay for each user is derived as [24], [28],

$$\begin{aligned} &D_{\text{Transport}} \\ &\quad = D_{\text{T-DU}} + D_{\text{T-CU}} \\ &\quad = \frac{1}{R_k - \lambda_k} + \sum_{m\in\mathcal{M}} \beta_{k,m} \frac{1}{r_m - \rho_m(L)}, \quad \forall k \in \mathcal{K}. \end{aligned} \tag{10}$$

We tend to allocate resources such that the user's delay requirements are met. This constraint can be mathematically represented as

$$\text{C8}: D_{\text{Total}} \le D_{\max}, \quad \forall k \in \mathcal{K}.$$

E. POWER CONSUMPTION MODEL

The total energy consumption in a predefined unit time, i.e., total power of our system, comes from the following components [29], [30]:

1) *The CU power:* The consumption power at the CU is proportional to the number of active VMs that can be dynamically turned on or off according to the system demands [23]. We consider the following relation for the number of active VMs (ν) as

$$\nu = \lceil \frac{\rho^{\text{FH}}(L)}{C^{\text{VM}}} \rceil, \tag{11}$$

where $\rho^{\text{FH}}(L) = \sum_{m\in\mathcal{M}} \rho_m$ is the total fronthaul traffic load and C^{VM} is the maximum capacity of each VM. If we consider the cost of each VM is ϕ, the total cost of the CU is

$$P_{\text{CU}} = \nu \times \phi. \tag{12}$$

2) *Power consumed by DUs:* Power consumption in each DU is from three main parts. First, processing power in DUs which depends on the level of FS. For simplicity, we consider a monotonic linear increasing function between DUs processing power and FS level, therefore,

$$P_{\text{proc}} = \sum_{m\in\mathcal{M}} b_m L P_0, \tag{13}$$

where P_0 is the normalized processing power of each DU. Second, transmit power of CU to DUs, i.e., $P_{\text{F}} = \sum_{n_{\text{F}}\in\mathcal{N}_{\text{F}}} p_{m,n_{\text{F}}}$ and the third is the circuit power consumption $P_C(b_m)$ of each DU, which is derived as [31]

$$P_{\text{C}}(b_m) = P_{\text{C0}} + b_m P_{\text{C1}}, \tag{14}$$

where P_{C0} and P_{C1} are the power consumption in sleep mode and the additional power of each DU in active mode, respectively. Therefore,

$$P_{\text{DU}} = P_{\text{proc}} + P_{\text{C}} + P_{\text{F}}. \tag{15}$$

3) *The power incurred by users:* Total receive power of each user from its dedicated DU, is equal to

$$P_{\text{User}} = \sum_{k\in\mathcal{K}} \sum_{n_{\text{A}}\in\mathcal{N}_{\text{A}}} \sum_{m\in\mathcal{M}} \alpha_{k,n_{\text{A}},m} p_{k,n_{\text{A}},m}. \tag{16}$$

Therefore, the total power of the system is sum powers of CU, DUs and users, i.e.,

$$P_{\text{Total}} = P_{\text{CU}} + P_{\text{DU}} + P_{\text{User}}. \tag{17}$$

Algorithm 1 Iterative Five-Step Solution

```
for $L = 1:8$ do
    Initialization;
    repeat
        Step 1: Access Power Allocation
        repeat
            Apply D.C. approximaition and find $P_{\mathrm{A}}$,
        until $P_A$ converge;
        Step 2: User Association
        Step 3: Power Allocation for DUs
        Step 4: Sub-carrier Allocation for DUs
        Step 5: Fronthaul Bandwidth Allocation
    until $P_A$ and $P_F$ converge;
    Compute Objective = $\frac{P\prime_{\mathrm{Total}}}{P_{\mathrm{O}}} + \frac{W\prime_{\mathrm{Total}}}{W_{\mathrm{O}}}$;
end
Select $L$ which minimize the objective.
```

III. PROBLEM FORMULATION AND PROPOSED ITERATIVE SOLUTION

A. PROBLEM FORMULATION

In our system model, we consider that the baseband processing functions can be flexibly split up between CU and DUs. The more functions centralized perform in CU, the greater the saving in the consumption power of the system. However, centralized function implementation increases the fronthaul links load and required fronthaul bandwidth. On the other hand, implementing more functions in DUs reduces the fronthaul load and requires less bandwidth but consume higher processing power. We remark this tradeoff and define our goal as the minimization of a weighted sum of the normalized system power and fronthaul bandwidth consumption under mentioned constraints in the previous section. The resource allocation problem can be formulated as

$$\min_{\mathbf{P},\boldsymbol{\alpha},\boldsymbol{\beta},L} \frac{P_{\mathrm{Total}}}{P_{\mathrm{O}}} + \frac{W_{\mathrm{Total}}}{W_{\mathrm{O}}}, \quad \text{subject to: C1-C8.} \tag{18}$$

where P_{O} and W_{O} are the normalization factors for total power and fronthaul bandwidth, respectively, $W_{\mathrm{Total}} = N_{\mathrm{F}} W_{\mathrm{F}}$ is total bandwidth of fronthaul links and $\mathbf{P} = [\mathbf{P}_{\mathrm{A}}, \mathbf{P}_{\mathrm{F}}]$ is the matrix of access and fronthaul power allocation vectors. In this problem, the users rate, i.e. R_k, is a non-convex function, $\alpha_{k,n_{\mathrm{A}},m}$, $\beta_{k,m}$, b_m, $\tau_{m,n_{\mathrm{F}}}$ are binary and L is an integer variable. Thus the problem is an NP-hard and non-convex. One approach to overcome this issue is dividing the problem into several sub-problems with a different time-scale, where is suited for the scenario with substantial different speed for variables change [32]–[34]. However, in this paper due to wireless fronthaul link, both access and fronthaul links variables have small-scale fading and we cannot separate their resource allocation problem in different timescales. We employ an iterative algorithm based on the block coordinate descent (BCD) method [35], where at each iteration, a single block of variables is optimized, while the remaining variables are fixed. The convergence of this method is guaranteed if the sub-problems are exactly solved to its unique optimal solution [35]. Although in [36] the general convergence analysis of successive convex approximation (SCA) method has been proven.

We divide the problem (18) into five sub-problems and solve them alternately. The solution of each sub-problem directly affects that of the next. The iterative procedure is continued until converge. We summarize the proposed iterative algorithm in Alg. 1. The detail of sub-problems are described in the following sub-sections.

B. ACCESS POWER ALLOCATION SUB-PROBLEM

By considering $\mathbf{P}_{\mathrm{A}}$ as the variable and all the other parameters constant, the problem convert as follows,

$$\begin{aligned} &\min_{\mathbf{P}_{\mathrm{A}}} \sum_{k\in\mathcal{K}} \sum_{n_{\mathrm{A}}\in\mathcal{N}_{\mathrm{A}}} \sum_{m\in\mathcal{M}} \beta_{k,m}\alpha_{k,n_{\mathrm{A}},m}p_{k,n_{\mathrm{A}},m}, \\ &\text{subject to: } \mathrm{C5}: \sum_{k\in\mathcal{K}} \sum_{n_{\mathrm{A}}\in\mathcal{N}_{\mathrm{A}}} p_{k,n_{\mathrm{A}},m} \le P_{\text{max-DU}}, \quad \forall m\in\mathcal{M}, \\ &\qquad \mathrm{C8}: \frac{1}{R_k - \lambda_k} \le D_{1,\max}, \quad \forall k\in\mathcal{K}, \end{aligned} \tag{19}$$

where $D_{1,\max}$ is a constant and derives as

$$D_{1,\max} = D_{\max} - \sum\nolimits_{m\in\mathcal{M}} \beta_{k,m}\left(\frac{1}{r_m - \rho_m(L)} + Ld_p\right), \quad \forall k\in\mathcal{K}, \tag{20}$$

and R_k is defined in (4). This problem is a non-convex due to the interference term in R_k. We adopt the following SCA approach by applying Difference-of-two-Concave-functions (D.C.) approximation [37] as

$$\log\Big(1 + \frac{p_{k,n_{\mathrm{A}},m}h_{k,n_{\mathrm{A}},m}}{\sigma^2 + \sum_{\substack{m'\in\mathcal{M},\\ m'\neq m}} \sum_{\substack{k'\in\mathcal{K}\\ k'\neq K}} p_{k\prime,n_{\mathrm{A}},m'}h_{k,n_{\mathrm{A}},m'}}\Big) = f_{k,n_{\mathrm{A}},m}(\mathbf{P}_{\mathrm{A}}) - g_{k,n_{\mathrm{A}},m}(\mathbf{P}_{\mathrm{A}}), \tag{21}$$

where

$$f_{k,n_{\mathrm{A}},m}(\mathbf{P}_{\mathrm{A}}) = \log(\sigma^2 + p_{k,n_{\mathrm{A}},m}h_{k,n_{\mathrm{A}},m} + \sum_{\substack{m'\in\mathcal{M},\\ m'\neq m}} \sum_{\substack{k'\in\mathcal{K}\\ k'\neq K}} p_{k',n_{\mathrm{A}},m'}h_{k,n_{\mathrm{A}},m'}), \tag{22}$$

and

$$g_{k,n_{\mathrm{A}},m}(\mathbf{P}_{\mathrm{A}}) = \log(\sigma^2 + \sum_{\substack{m'\in\mathcal{M},\\ m\prime\neq m}} \sum_{\substack{k'\in\mathcal{K}\\ k\prime\neq K}} p_{k',n_{\mathrm{A}},m'}h_{k,n_{\mathrm{A}},m'}). \tag{23}$$

According to D.C., we employ the following approximation:

$$g(\mathbf{P}_{\mathrm{A}}(t_1)) \approx g(\mathbf{P}_{\mathrm{A}}[t_1-1]) + \nabla g^T(\mathbf{P}_{\mathrm{A}}[t_1-1])(\mathbf{P}_{\mathrm{A}} - \mathbf{P}_{\mathrm{A}}[t_1-1]), \tag{24}$$

where $\mathbf{P}_{\mathrm{A}}[t_1-1]$ is specified from iteration $t_1 - 1 \geq 0$ and $\nabla g(\mathbf{P}_{\mathrm{A}}[t_1-1])$ is the derivation of $g(\mathbf{P}_{\mathrm{A}}[t_1-1])$. By applying D.C. approximation, the problem of $\mathbf{P}_{\mathrm{A}}$ allocation in each step, converts to a convex form and can be solved by CVX. The iterative algorithm stops when $\mathbf{P}_{\mathrm{A}}$ converge.

C. USER ASSOCIATION SUB-PROBLEM

For given values of $\mathbf{P}_{\mathrm{F}}, \boldsymbol{\tau}, W_{\mathrm{F}}, L$, and obtained $\mathbf{P}_{\mathrm{A}}$ from step 1, the optimization problem is formulated as

$$\begin{aligned}
&\min_{\boldsymbol{\beta},b,\boldsymbol{\alpha}} \sum_{k\in\mathcal{K}}\sum_{n_{\mathrm{A}}\in\mathcal{N}_{\mathrm{A}}}\sum_{m\in\mathcal{M}} \beta_{k,m}\alpha_{k,n_{\mathrm{A}},m}p_{k,n_{\mathrm{A}},m} + \sum_{m\in\mathcal{M}} b_m P_{FC}\\
&\text{subject to}:\\
&\qquad \mathrm{C2}: \sum_{k\in\mathcal{K}} \alpha_{k,n_{\mathrm{A}},m} \leq 1 \quad \forall n_{\mathrm{A}} \in \mathcal{N}_{\mathrm{A}}, \forall m \in \mathcal{M},\\
&\qquad \mathrm{C3}: \sum_{m\in\mathcal{M}} \beta_{k,m}(t) \leq 1, \quad \forall k \in \mathcal{K},\\
&\qquad \mathrm{C4}: \beta_{k,m} \leq b_m, \quad \forall k \in \mathcal{K}, \ \forall m \in \mathcal{M},\\
&\qquad \mathrm{C8}: \frac{1}{R_k - \lambda_k} \leq D_{1,\max}, \quad \forall k \in \mathcal{K}, \qquad (25)
\end{aligned}$$

where

$$P_{FC} = LP_{\mathrm{proc}} + \sum_{n_{\mathrm{F}}\in\mathcal{N}_{\mathrm{F}}} \tau_{m,n_{\mathrm{F}}}p_{m,n_{\mathrm{F}}} + P_{C1} - P_{C0}, \qquad (26)$$

and $D_{1,\max}$ is defined in (20). The problem (25) is a Mixed Integer Non-Linear Programming (MINLP) and NP-hard, due to the quadratic objective function. To simplify it, we replace the product of two binary variables $\alpha_{k,n_{\mathrm{A}},m}$ and $\beta_{k,m}$ with a new binary variable $\zeta_{k,n_{\mathrm{A}},m}$ and add following new constraints [38]:

$$\begin{aligned}
\mathrm{C0}_1 &: \zeta_{k,n_{\mathrm{A}},m} \leq \alpha_{k,n_{\mathrm{A}},m},\\
\mathrm{C0}_2 &: \zeta_{k,n_{\mathrm{A}},m} \leq \beta_{k,m},\\
\mathrm{C0}_3 &: \zeta_{k,n_a,m} \geq \alpha_{k,n_{\mathrm{A}},m} + \beta_{k,m} - 1.
\end{aligned}$$

Then (25) transform into a linear integer programming problem and can be solved by MOSEK solver of CVX which utilizes the interior-point method [39].

D. POWER ALLOCATION FOR DUS SUB-PROBLEM

For a given values of $\boldsymbol{\tau}, W_{\mathrm{F}}, L$, and obtained values for $\mathbf{P}_{\mathrm{A}}, \boldsymbol{\alpha}, \boldsymbol{\beta}$, and b in steps 1-2, the optimization problem is transformed to

$$\begin{aligned}
&\min_{\mathbf{P}_{\mathrm{F}}} \sum_{m\in\mathcal{M}}\sum_{n_{\mathrm{F}}\in\mathcal{N}_{\mathrm{F}}} b_m\tau_{m,n_{\mathrm{F}}}p_{m,n_{\mathrm{F}}}\\
&\quad \mathrm{C7}: \sum_{m\in\mathcal{M}}\sum_{n_{\mathrm{F}}\in\mathcal{N}_{\mathrm{F}}} \tau_{m,n_{\mathrm{F}}}p_{m,n_{\mathrm{F}}} \leq P_{\text{max-CU}}\\
&\quad \mathrm{C8}: \sum_{m\in\mathcal{M}} \beta_{k,m}\frac{1}{r_m - \rho_m} \leq D_{2,\max}, \quad \forall k \in \mathcal{K}, \qquad (27)
\end{aligned}$$

where r_m is defined in (6) and

$$D_{2,\max} = D_{\max} - \sum_{m\in\mathcal{M}} \beta_{k,m}\left(\frac{1}{R_k - \lambda_k} + Ld_p\right). \qquad (28)$$

This problem is convex and the optimal value of $\mathbf{P}_{\mathrm{F}}$ can be derived by CVX.

E. DUS SUB-CARRIER ALLOCATION SUB-PROBLEM

For given values of W_{F}, L, and obtained values for $\mathbf{P}_{\mathrm{A}}, \boldsymbol{\alpha}, \boldsymbol{\beta}, b$, and $\mathbf{P}_{\mathrm{F}}$ in steps 1-3 we have

$$\begin{aligned}
&\min_{\boldsymbol{\tau}} \sum_{m\in\mathcal{M}}\sum_{n_{\mathrm{F}}\in\mathcal{N}_{\mathrm{F}}} b_m\tau_{m,n_{\mathrm{F}}}p_{m,n_{\mathrm{F}}}\\
&\text{subject to}:\\
&\qquad \mathrm{C6}: \tau_{m,n_{\mathrm{F}}} \leq 1, \quad \forall m \in \mathcal{M} \text{ and } n_{\mathrm{F}} \in \mathcal{N}_{\mathrm{F}},\\
&\qquad \mathrm{C8}: \sum_{m\in\mathcal{M}} \beta_{k,m}\frac{1}{r_m - \rho_m} \leq D_{2,\max}, \quad \forall k \in \mathcal{K}, \qquad (29)
\end{aligned}$$

where $D_{2,\max}$ is defined in (28). This problem is also convex and can be solved by CVX.

F. FRONTHAUL BANDWIDTH ALLOCATION SUB-PROBLEM

For a given value of L and derived values for other variables in previous steps, the fronthaul bandwidth allocation sub-problem can be expressed as

$$\begin{aligned}
&\min_{W_{\mathrm{F}}} \frac{N_{\mathrm{F}}W_{\mathrm{F}}}{W_{\mathrm{O}}}\\
&\text{subject to}: \ \mathrm{C8}: \sum_{m\in\mathcal{M}} \beta_{k,m}\frac{1}{r_m - \rho_m} \leq D_{2,\max}, \quad \forall k \in \mathcal{K}. \qquad (30)
\end{aligned}$$

By considering (6), this problem is linear relative to W_{F} and can simply be solved by CVX.

The overall optimization problem is iteratively solved until $P_{\mathbf{F}}$ and $P_{\mathbf{A}}$ converge, i.e, the variation from the previous iteration is less than a predefined $\epsilon \ll 1$. The problem is solved for all $L \in \{1, 2, \ldots, 8\}$ and the best values of the FS level for all DUs which minimize the objective function is selected. Resource allocation calculations perform in CU and due to the existence of a strong processor in it, we neglect its processing delay of this procedure.

IV. COMPLEXITY AND CONVERGENCE ANALYSIS

In this section, we investigate the computational complexity and convergence analysis of the proposed algorithm. The first sub-problem is solved based on D.C. approximation method and the complexity of it is $O(\log(1/\epsilon_{\mathrm{dc}}))$, where ϵ_{dc} is the stopping criteria [40]. After applying D.C. approximation, the problem can solved by Lagrangian method in CVX, which have the complexity $O(n_v/\epsilon_{\mathrm{sub}}^2)$ where $1/\epsilon_{\mathrm{sub}}^2$ is the number of iterations of sub-gradient method to find a ϵ_{sub} sub-optimal point. Therefore, the total complexity of access power allocation sub-problem is $O(n_v/(\epsilon_{\mathrm{sub}}^2\epsilon_{\mathrm{dc}}))$ [41]. The second sub-problem of the proposed algorithm is solved based on the linear programming primal-dual interior-point approach, and in the worst case require $O(\sqrt{n_v}n_s)$ iterations to approach the feasible solution, where n_v is the number of variables and n_s is the size of the problem data [42]. The third sub-problem is convex and can be solved by the Lagrangian method in CVX. The fourth and fifth sub-problems of the proposed algorithm are according to the linear programming. Thus the complexity of the proposed algorithm only grows

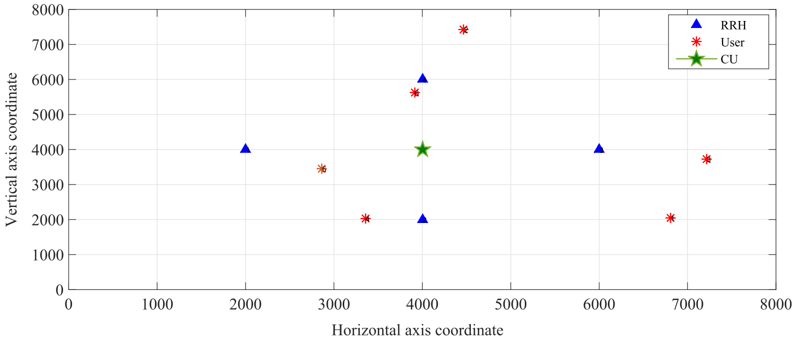

FIGURE 4. An example realization of network setup in simulation results.

polynomially with the number of variables and is notably better in contrast to the direct search methods with exponential complexities.

Regarding the convergence, the proposed algorithm is based on the BCD method, which can be applied to minimize non-convex functions with certain separability and regularity properties. According to that, at each iteration, one group of variables is optimized while the remaining of them are assumed to be fixed. The results of the previous iterations are applied to the optimized current iteration. The convergence of the BCD method is guaranteed, if the sub-problem of each iteration is exactly solved to its unique optimal solution [35] or even updated by SCA [36]. Therefore, the convergence of the proposed algorithm is established. However, it converges to local optimum and may not be coincident global optimum.

V. SIMULATION RESULTS

A. NETWORK SETUP

In this section, the simulation results for different system parameters are presented to evaluate the performance of the proposed algorithm via MATLAB Monte Carlo simulations through 100 network realizations. We consider a CU at the center of a 8 km square coverage area whose distance is 2 Km from $M = 4$ DUs. An example realization of the considered network topology for $K = 6$ is displayed in Fig. 4. The channel power gains are based on the path loss and Rayleigh fading for both access and fronthaul links, i.e. $h_{k,n_{\text{A}},m} = \xi_{k,n_{\text{A}},m} d_{k,m}^{-\theta}$ and $g_{m,n_{\text{F}}} = \zeta_{m,n_{\text{F}}} \delta_m^{-\theta}$, where $\xi_{k,n_{\text{A}},m}$ and $\zeta_{m,n_{\text{F}}}$ are random variables which generated by Rayleigh distribution, $d_{k,m}$ is the distance of user k and DU m, δ_m is the distance of DU m and CU, and $\theta = 3$ is the path-loss exponent. The noise power is set to -174 dBm/Hz. For each DU and CU, the maximum power are set as $P_{\text{max-DU}} = 28$dBm and $P_{\text{max-CU}} = 35$dBm, respectively. The mean arrival rate of each user is assumed $\lambda_k = 10^5$bit/s. The bandwidth of each sub-carrier in access links is $W_{\text{A}} = 700$kHz. All the considered parameters are summarized in Table 1.

TABLE 1. Considered parameters in numerical results.

Symbol	Quantity	Value
σ^2	Noise power	-174 dBm/Hz
W_{O}	Bandwidth normalization factor	10^6
P_{O}	Power normalization factor	10
λ_k	Mean arrival rate	10^5
W_{A}	Access sub-carrier bandwidth	700 kHz
$P_{\text{max-DU}}$	Max. Transmit power of DUs	28 dBm
$P_{\text{max-CU}}$	Max. Transmit power of CU	35 dBm
K	Number of users	6
M	Number of DU	4
N_{A}	Number of Access sub-carriers	8
N_{F}	Number of Fronthaul sub-carriers	6
θ	Path-loss exponent	3
d_p	Normalized time for process in DUs	20 μs
P_{C0}	DU power consumption in sleep mode	3 dBm
P_{C1}	DU power consumption in active mode	17 dBm
L_0	Constant normalization for fronthaul rate	10
P_{proc}	Normalized processing power of DUs	23 dBm
Φ	Normalized power of each VM	3 dBm
C_{VM}	Capacity of each VM	0.1MHz

B. CONVERGENCE ANALYSIS

Fig. 5 illustrates the convergence behaviour of the proposed algorithm for different FS level with $D_{\max} = 1$ms. The sum of the normalized power and fronthaul bandwidth with the normalization factors $P_{\text{O}} = 10$ and $W_{\text{O}} = 10^5$, respectively, is considered as the objective function. Numerical results confirm that our proposed five steps algorithm converges in about 20 iterations. Furthermore, Fig. 5 shows that for considered setup, $L = 4$ gives the minimum objective value, which is about 40% better in contrast to the traditional C-RAN strategy i.e. $L = 1$.

C. EFFECTS OF NETWORK PARAMETERS

In Fig. 6 we evaluate the effect of FS level on the delay components. As can be seen, by increasing the FS level, processing time in DUs increases and the transporting delay of CU reduces. This is due to that implementing more functions in DUs, increases processing time, but reduces the fronthaul traffic load and the CU transporting delay $D_{\text{T-CU}}$. For example, as FS level increase from 1 to 2, processing delay grows 50% (0.02ms) and $D_{\text{T-CU}}$ decrease about 0.03ms which is 0.05%. We also see that $D_{\text{T-DU}}$ is almost constant for different FS level. For our setup, the total delay has a slight decrease by increasing the FS level. For $D_{\max} = 1$ms, D_{Total} decrease from 0.65ms in FS level 1 to 0.56ms in FS level 8.

The effect of FS level on all the considered power components is investigated in Fig. 7 for $D_{\max} = 1$ms. As we

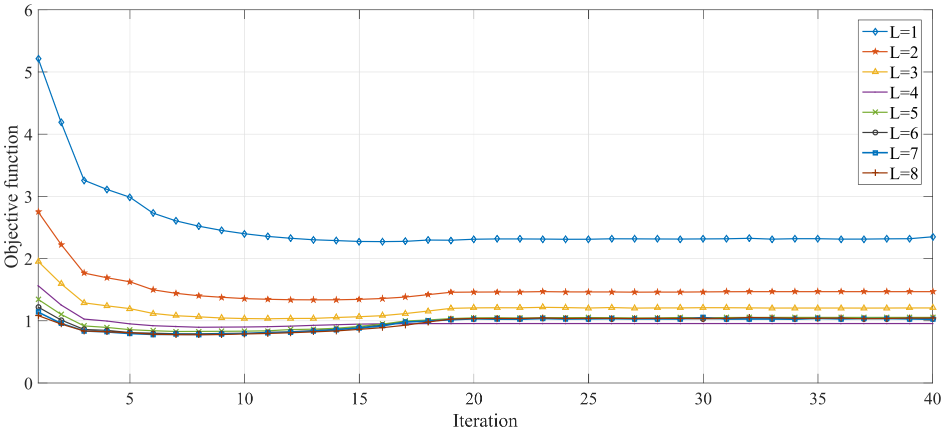

FIGURE 5. The convergence in terms of of objective function over the number of iterations.

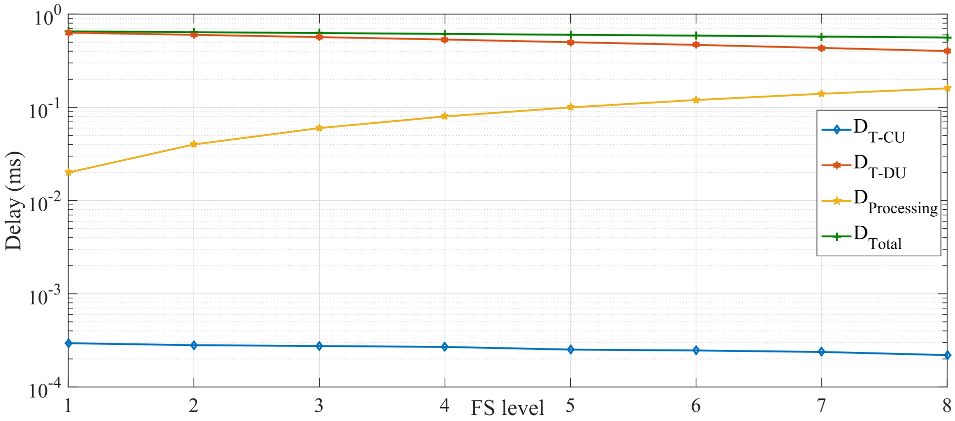

FIGURE 6. Delay components versus FS level for $D_{\max} = 1$ms.

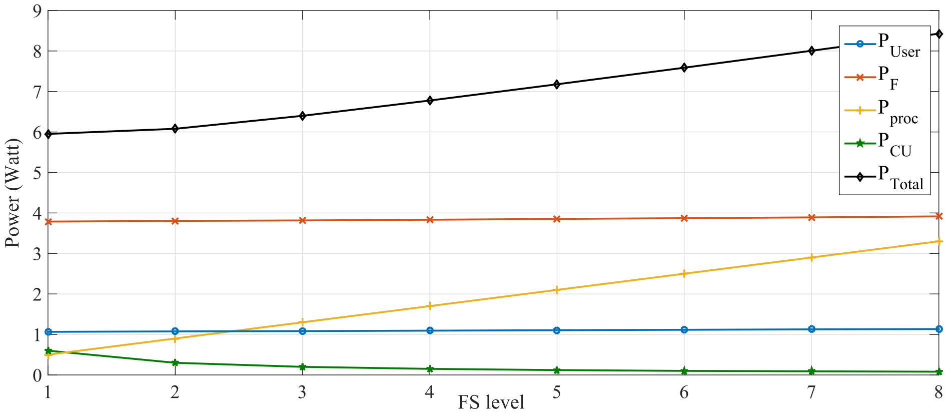

FIGURE 7. Power components versus FS level for $D_{\max} = 1$ms.

expected, by increasing the FS level, i.e. more functions implement in DUs, processing power in DUs (P_{proc}) increases while in CU (P_{CU}) decreases. In other words, for a high FS level, the processing power of CU is low and negligible compared to other power components. Also, we can see that the transmit power of users and the DUs are almost constant in FS level. Overall total power increase from 5.9516 to 8.4279 watt (about 40%) when FS level rises from 1 to 8.

Fig. 8 illustrates the impact of the FS level on the required fronthaul bandwidth for different delay constraints. As FS level increases, the traffic load on fronthaul decreases and hence less fronthaul bandwidth is required. For strict delay

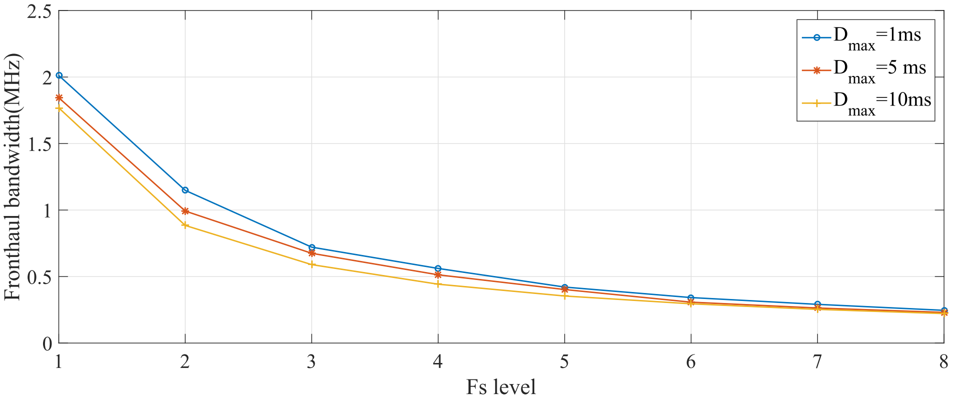

FIGURE 8. Fronthaul bandwidth versus FS level for different $D_{\max}$.

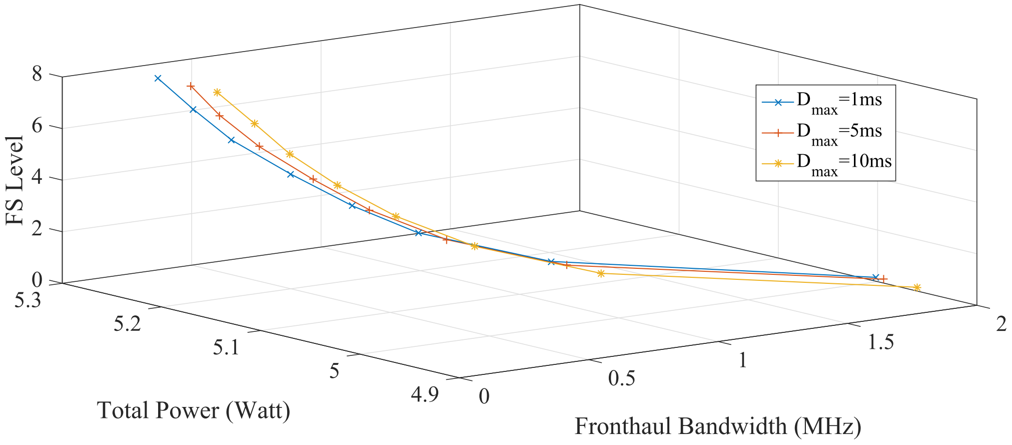

FIGURE 9. Fronthaul bandwidth and total power versus FS level for different $D_{\max}$.

constraints, we need more fronthaul bandwidth to meet the delay requirement. For example, in FS level 2, as the $D_{\max}$ increase from 1 ms to 10 ms, the required fronthaul bandwidth decrease from 1.15 MHz to 0.9 MHz, which is about 20% degradation in it.

By combining Fig. 7 and Fig. 8, we derive a 3D plot of total power versus fronthaul bandwidth for different FS levels in Fig. 9. We can see from this figure that increasing the delay threshold cause to decrease in fronthaul power and required fronthaul bandwidth. Moreover, for considered simulation setup, in different delay thresholds, the best function splitting level which minimizes objective function is fixed. However, it can be changed by varying the normalization factors of power and bandwidth.

Fig. 10 illustrates the impact of the number of users on the objective function (i.e., the sum of normalized total power and fronthaul bandwidth consumption) for both flexible FS and traditional C-RAN scenarios ($L = 1$) when $D_{\max} = 1$ms. Increasing the number of users leads to an increase in total power and fronthaul bandwidth, and consequently higher objective function. Also, we see that the gain of flexible FS increases for the higher number of users. For example, when number of users grows from K = 6 to K = 16 the gain of FS in contrast to the traditional C-RAN grows from 39.26% to 42.65%, which implies that flexible FS becomes more important with a larger number of users.

D. SUMMARY OF SIMULATION RESULTS

The main simulation results of this paper are summarized as follows:

1) For considered setup and different FS level (eight options introduced by 3GPP), we derive the total utility function (i.e., the sum of normalized transmit power and required bandwidth) which are directly related to the cost of the network. For high levels of FS, the transmitting data between DUs and CU is decreasing, thus less fronthaul bandwidth is required. On the other hand, in high FS levels, more functions are implemented locally in DUs, which leads to an increase in the total power of the system. In considered total utility, the normalization factors balance the weight of power and bandwidth consumption. By flexibly selection the FS level which minimizes the total utility function, in our system setup, we derive 40% gain in contrast to traditional C-RAN (L = 1).

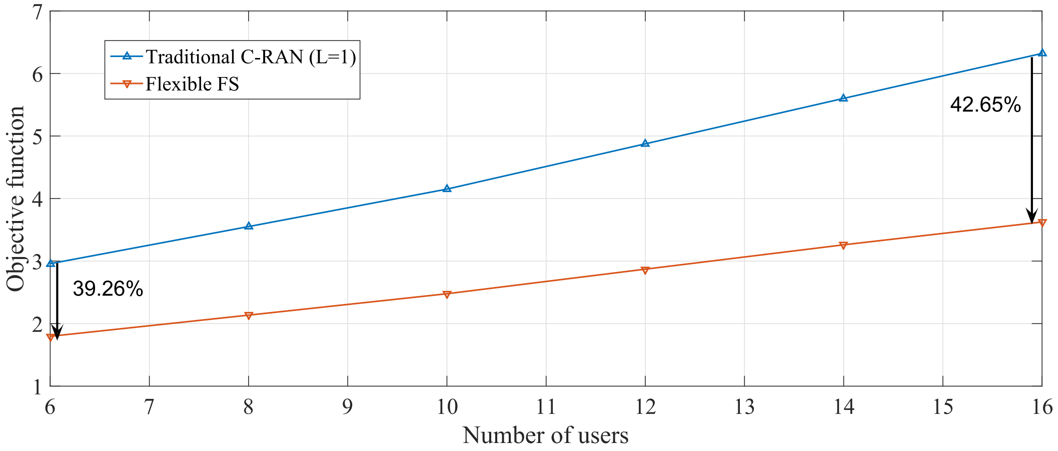

FIGURE 10. **Objective function versus number of users.**

2) Implementing more functions in DUs (high FS level), increase total processing delay, but reduces the transporting delay of CUs (due to lower transferred data between DU and CU). Overall total delay show about 14% decrease from FS 1 to 8.
3) For strict delay constraint, more fronthaul bandwidth is needed. However, by increasing the FS level, we can save bandwidth with the cost of power. Flexible adjusting FS level according to the system constraint, (delay, power, ...) is a good solution that can save system resources and the achievable gain of it increases with the number of users.

VI. CONCLUSION

In this paper, we propose a flexible function split between CU and DUs in downlink C-RAN for delay critical applications. We incorporate total delay components consist of processing, transmission and queueing. We present a double layer queuing model, wherein the first layer is for the aggregated transmit data of each DU and the second layer is for the data of each user. Tack into account sum of the normalized both fronthaul bandwidth and total power of the system as an objective function, we formulate a joint fronthaul and access resource allocation problem with delay constraint. Due to the non-convex nature of the proposed problem, an iterative algorithm based on the BCD method is applied. We solve the problem for all $L = 1, 2, \ldots, 8$ and select the best FS level for all DUs which minimize the objective function. Simulation results demonstrate that delay constraint has a significant impact on the required fronthaul bandwidth and power consumption. As the delay threshold increase from 1ms to 10ms, we can see up to 20% degradation in required fronthaul bandwidth. We also find that the flexible selection of function split level can derive about 40% gain in terms of our considered objective function and is increasing with the number of users.

REFERENCES

[1] J. Tang, R. Wen, T. Q. S. Quek, and M. Peng, "Fully exploiting cloud computing to achieve a green and flexible C-RAN," *IEEE Commun. Mag.*, vol. 55, no. 11, pp. 40–46, Nov. 2017, doi: 10.1109/MCOM.2017.1600922.

[2] I. T. Haque and N. Abu-Ghazaleh, "Wireless software defined networking: A survey and taxonomy," *IEEE Commun. Surveys Tutr.*, vol. 18, no. 4, pp. 2713–2737, 4th Quart., 2016, doi: 10.1109/COMST.2016.2571118.

[3] M. Baghani, S. Parsaeefard, and T. Le-Ngoc, "Multi-objective resource allocation in density-aware design of C-RAN in 5G," *IEEE Access*, vol. 6, pp. 45177–45190, 2018, doi: 10.1109/ACCESS.2018.2861909.

[4] U. Dötsch, M. Doll, H.-P. Mayer, F. Schaich, J. Segel, and P. Sehier, "Quantitative analysis of split base station processing and determination of advantageous architectures for LTE," *Bell Labs Tech. J.*, vol. 18, no. 1, pp. 105–128, Jun. 2013, doi: 10.1002/bltj.21595.

[5] L. M. P. Larsen, A. Checko, and H. L. Christiansen, "A survey of the functional splits proposed for 5G mobile crosshaul networks," *IEEE Commun. Surv. Tuts.*, vol. 21, no. 1, pp. 146–172, 1st Quart., 2019, doi: 10.1109/COMST.2018.2868805.

[6] *Study on New Radio Access Technology: Radio Access Architecture and Interfaces*, 3GPP, document 3GPP TR 38.801 V14.0.0 (2017-03), 2017.

[7] J. Bartelt, P. Rost, D. Wubben, J. Lessmann, B. Melis, and G. Fettweis, "Fronthaul and backhaul requirements of flexibly centralized radio access networks," *IEEE Wireless Commun.*, vol. 22, no. 5, pp. 105–111, Oct. 2015, doi: 10.1109/MWC.2015.7306544.

[8] J. Bartelt, N. Vucic, D. Camps-Mur, E. Garcia-Villegas, I. Demirkol, A. Fehske, M. Grieger, A. Tzanakaki, J. Gutiérrez, E. Grass, G. Lyberopoulos, and G. Fettweis, "5G transport network requirements for the next generation fronthaul interface," *EURASIP J. Wireless Commun. Netw.*, vol. 2017, no. 1, p. 89, Dec. 2017, doi: 10.1186/s13638-017-0874-7.

[9] M. Peng, Y. Sun, X. Li, Z. Mao, and C. Wang, "Recent advances in cloud radio access networks: System architectures, key techniques, and open issues," *IEEE Commun. Surveys Tutr.*, vol. 18, no. 3, pp. 2282–2308, 3rd Quart., 2016, doi: 10.1109/COMST.2016.2548658.

[10] C.-Y. Chang, N. Nikaein, R. Knopp, T. Spyropoulos, and S. S. Kumar, "FlexCRAN: A flexible functional split framework over ethernet fronthaul in Cloud-RAN," in *Proc. IEEE Int. Conf. Commun.(ICC)*, Paris, France, May 2017, pp. 1–7.

[11] A. Garcia-Saavedra, X. Costa-Perez, D. J. Leith, and G. Iosifidis, "Fluidran: Optimized vRAN/MEC orchestration," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Honolulu, HI, USA, Apr. 2018, pp. 2366–2374.

[12] D. Harutyunyan and R. Riggio, "Flex5G: Flexible functional split in 5G networks," *IEEE Trans. Netw. Serv. Manage.*, vol. 15, no. 3, pp. 961–975, Sep. 2018, doi: 10.1109/TNSM.2018.2853707.

[13] Y. Zhou, J. Li, Y. Shi, and V. W. S. Wong, "Flexible functional split design for downlink C-RAN with capacity-constrained fronthaul," *IEEE Trans. Veh. Technol.*, vol. 68, no. 6, pp. 6050–6063, Jun. 2019, doi: 10.1109/TVT.2019.2911934.

[14] S. Matoussi, I. Fajjari, N. Aitsaadi, R. Langar, and S. Costanzo, "Joint functional split and resource allocation in 5G cloud-RAN," in *Proc. IEEE Int. Conf. Commun.(ICC)*, Shanghai, China, May 2019, pp. 1–7.

[15] G. O. Pérez, J. A. Hernández, and D. Larrabeiti, "Fronthaul network modeling and dimensioning meeting ultra-low latency requirements for 5G," *J. Opt. Commun. Netw.*, vol. 10, no. 6, p. 573–581, Jun. 2018, doi: 10.1364/JOCN.10.000573.

[16] L. M. P. Larsen, M. S. Berger, and H. L. Christiansen, "Fronthaul for Cloud-RAN enabling network slicing in 5G mobile networks," *IEEE Wireless Commun. Lett.*, vol. 2018, pp. 573–581, Jul. 2018, doi: 10.1155/2018/4860212.

[17] A. Checko, A. P. Avramova, M. S. Berger, and H. L. Christiansen, "Evaluating C-RAN fronthaul functional splits in terms of network level energy and cost savings," *J. Commun. Netw.*, vol. 18, no. 2, pp. 162–172, Apr. 2016, doi: 10.1109/JCN.2016.000025.

[18] J. Kant Chaudhary, J. Francis, A. Noll Barreto, and G. Fettweis, "Latency in the uplink of massive MIMO CRAN with packetized fronthaul: Modeling and analysis," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Marrakech, Morocco, Apr. 2019, pp. 1–7.

[19] A. Alabbasi, X. Wang, and C. Cavdar, "Optimal processing allocation to minimize energy and bandwidth consumption in hybrid CRAN," *IEEE Trans. Green Commun. Netw.*, vol. 2, no. 2, pp. 545–555, Jun. 2018, doi: 10.1109/TGCN.2018.2802419.

[20] C.-Y. Chang, R. Schiavi, N. Nikaein, T. Spyropoulos, and C. Bonnet, "Impact of packetization and functional split on C-RAN fronthaul performance," in *Proc. IEEE Int. Conf. Commun.(ICC)*, Kuala Lumpur, Malaysia, May 2016, pp. 1–7.

[21] J. Yusupov, A. Ksentini, G. Marchetto, and R. Sisto, "Multi-objective function splitting and placement of network slices in 5G mobile networks," in *Proc. IEEE Conf. Standards for Commun. Netw. (CSCN)*, Paris, France, Oct. 2018, pp. 1–6.

[22] M. Peng, C. Wang, V. Lau, and H. V. Poor, "Fronthaul-constrained cloud radio access networks: Insights and challenges," *IEEE Wireless Commun.*, vol. 22, no. 2, pp. 152–160, Apr. 2015, doi: 10.1109/MWC.2015.7096298.

[23] D. Bhamare, A. Erbad, R. Jain, M. Zolanvari, and M. Samaka, "Efficient virtual network function placement strategies for Cloud Radio Access Networks," *Comput. Commun.*, vol. 127, pp. 50–60, Sep. 2018, doi: 10.1016/j.comcom.2018.05.004.

[24] D. Bertsekas and R. Gallager, "Delay models in data networks," in *Data Networks*, 2nd ed. Upper Saddle River, NJ, USA: Prentice-Hall, 1992, ch. 2, pp. 15–64.

[25] C. A. Garcia-Perez and P. Merino, "Enabling low latency services on LTE networks," in *Proc. IEEE 1st Int. Workshops Found. Appl. Self Syst. (FASW)*, Sep. 2016, pp. 248–255.

[26] I. Parvez, A. Rahmati, I. Guvenc, A. I. Sarwat, and H. Dai, "A survey on low latency towards 5G: RAN, core network and caching solutions," *IEEE Commun. Surveys Tutr.*, vol. 20, no. 4, pp. 3098–3130, 4th Quart., 2018, doi: 10.1109/COMST.2018.2841349.

[27] M. Tohidi, H. Bakhshi, and S. Parsaeefard, "Joint uplink and downlink delay-aware resource allocation in C-RAN," *Trans. Emerg. Telecommun.*, Dec. 2019, Art. no. e3778, doi: 10.1002/ett.3778.

[28] J. Tang, W. P. Tay, T. Q. S. Quek, and B. Liang, "System cost minimization in cloud RAN with limited fronthaul capacity," *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 3371–3384, May 2017, doi: 10.1109/TWC.2017.2682079.

[29] A. Younis, T. X. Tran, and D. Pompili, "Bandwidth and energy-aware resource allocation for cloud radio access networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 10, pp. 6487–6500, Oct. 2018, doi: 10.1109/TWC.2018.2860008.

[30] A. Younis, T. Tran, and D. Pompili, "Energy-efficient resource allocation in C-RANs with capacity-limited fronthaul," *IEEE Trans. Mobile Comput.*, to be published, 10.1109/TMC.2019.2942597.

[31] M. Sinaie, A. Zappone, E. A. Jorswieck, and P. Azmi, "A novel power consumption model for effective energy efficiency in wireless networks," *IEEE Wireless Commun. Lett.*, vol. 5, no. 2, pp. 152–155, Apr. 2016, doi: 10.1109/LWC.2015.2512259.

[32] J. Tang, B. Shim, and T. Q. S. Quek, "Service multiplexing and revenue maximization in sliced C-RAN incorporated with URLLC and multicast eMBB," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 4, pp. 881–895, Apr. 2019, doi: 10.1109/JSAC.2019.2898745.

[33] M. Wang, N. Karakoc, L. Ferrari, P. Shantharama, A. S. Thyagaturu, M. Reisslein, and A. Scaglione, "A multi-layer multi-timescale network utility maximization framework for the SDN-based layback architecture enabling wireless backhaul resource sharing," *Electronics*, vol. 8, no. 9, p. 937, Aug. 2019, doi: 10.3390/electronics8090937.

[34] W. Xia, T. Q. S. Quek, J. Zhang, S. Jin, and H. Zhu, "Programmable Hierarchical C-RAN: From Task Scheduling to Resource Allocation," *IEEE Trans. Wireless Commun.*, vol. 18, no. 3, pp. 2003–2016, Mar. 2019, doi: 10.1109/TWC.2019.2901684.

[35] P. Tseng, "Convergence of a block coordinate descent method for nondifferentiable minimization," *J. Optim. Theory Appl.*, vol. 109, no. 3, pp. 475–494, Jun. 2001, doi: 10.1023/A:101750170.

[36] M. Razaviyayn, M. Hong, and Z.-Q. Luo, "A unified convergence analysis of block successive minimization methods for nonsmooth optimization," *SIAM J. Optim.*, vol. 23, no. 2, pp. 1126–1153, Jan. 2013, doi: 10.1137/120891009.

[37] D. T. Ngo, S. Khakurel, and T. Le-Ngoc, "Joint subchannel assignment and power allocation for OFDMA femtocell networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 1, pp. 342–355, Jan. 2014, doi: 10.1109/TWC.2013.111313.130645.

[38] M. Y. Lyazidi, N. Aitsaadi, and R. Langar, "Dynamic resource allocation for Cloud-RAN in LTE with real-time BBU/RRH assignment," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Kuala Lumpur, Malaysia, May 2016, pp. 1–6.

[39] M. Grant and S. Boyd. (Mar. 2017). *CVX: MATLAB Software for Disciplined Convex Programming*. R Package Version 2.1. [Online]. Available: http://cvxr.com/cvx

[40] A. L. Yuille and A. Rangarajan, "The concave-convex procedure," *Neural Comput.*, vol. 15, no. 4, pp. 915–936, Apr. 2003, doi: 10.1162/08997660360581958.

[41] M. Baghani, S. Parsaeefard, M. Derakhshaniy, and W. Saad, "Dynamic non-orthogonal multiple access (NOMA) and orthogonal multiple access (OMA) in 5G wireless networks," *IEEE Trans. Commun.*, to be published, doi: 10.1109/TCOMM.2019.2919547.

[42] S. J. Wright, *Primal-Dual Interior-Point Methods*. Philadelphia, PA, USA: SIAM, 1997.

MASUMEHSADAT TOHIDI received the B.S. degree in electrical engineering from Kashan University, Kashan, Iran, in 2010, and the M.S. degree in communication engineering from Tarbiat Modares University, Tehran, Iran, in 2012. She is currently pursuing the Ph.D. degree in communication engineering with Shahed University. Her current research interest includes resource allocation in delay critical application.

HAMIDREZA BAKHSHI was born in Tehran, Iran, in April 1971. He received the B.Sc. degree in electrical engineering from the University of Tehran, in 1992, and the M.Sc. and Ph.D. degrees in electrical engineering from Tarbiat Modares University, Iran, in 1995 and 2001, respectively. He has been an Associate Professor of electrical engineering with Shahed University, Tehran, since 2010. His areas of research include wireless communications, multiuser detection, channel estimation, cognitive radio, and smart antennas.

SAEEDEH PARSAEEFARD (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees from the Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran, in 2003 and 2006, respectively, and the Ph.D. degree in electrical and computer engineering from Tarbiat Modares University, Tehran, in 2012. She was a Postdoctoral Research Fellow with the Telecommunication and Signal Processing Laboratory, Department of Electrical and Computer Engineering, McGill University, Canada. From 2010 to 2011, she was a Visiting Ph.D. Student with the Department of Electrical Engineering, University of California at Los Angeles, CA, USA. She is currently a Faculty Member with the Iran Telecommunication Research Center and a Visiting Faculty Member with the University of Toronto. Her current research interests include the resource management in software-defined networking, the Internet of Things, and the fifth generation of wireless networks, as well as applications of robust optimization theory and game theory on the resource allocation and management in wireless networks. She received the IEEE Iran Section Women in Engineering (WIE) Award, in 2018.